

Towards a 'knowledge nation'

With a federal election due in Australia this year, the two main political parties are vying over how best to stimulate the growth of biotechnology. But the quest for a knowledge-intense economy requires long-term investment and management.

Australia has missed many opportunities to commercialize potentially valuable research discoveries. This has been a result of both inadequate financial backing by government and industry and a lack of commercial expertise and entrepreneurial spirit within the research community. Much of this is changing, and the biotechnology industry is now buoyed by a national sense of optimism. So what should be done to help it on its way?

Australia's desire to shift towards a knowledge-intense economy, along similar lines to Ireland and Canada, has been signalled by various federal government initiatives. Launched in January this year, the federal innovation strategy 'Backing Australia's Ability' has pledged A\$2.9 billion (US\$1.5 billion) over five years to stimulate the commercialization of research discoveries. It includes funds to help bring biotechnology initiatives to the proof-of-concept stage. It is also giving A\$155 million over five years to the Major National Research Facilities programme to cover half the project costs — the rest to come from universities and industry — of creating large-scale specialized research facilities. Contenders for this hotly contested programme include proposals for the nation's first synchrotron.

But the Conservative federal government's investment is unlikely to propel Australia to the front line of the global biotechnology market. Most of the A\$2.9 billion is slated for the final years of the programme. Indeed, Kim Beazley, the leader of the Labor opposition party, claims that the innovation strategy does little more than put back some of the A\$5 billion stripped from government spending on innovation and education over the past five years.

States' initiatives

Fortunately, biotechnology has been supported by the governments of Australia's states, which fund infrastructure and local research initiatives. The competition between the states, particularly along the eastern coast, has accelerated the growth of biotechnology, with each state pushing to establish its capital city as the premier centre for biotechnology research and investment. Recent announcements have underscored the states' goals to set up biotechnology 'hubs' or 'clusters', which draw together research institutes, teaching hospitals and start-up companies in local precincts to fuel the translation of basic science into commercial products and clinical applications. Although this fosters biotech growth at the local level, a national framework to coordinate such efforts would harness the benefits that stem from competition while avoiding redundancy of effort and maximizing the efficient use of national resources.

Although parts of the Australian public sector are abuzz with bio-innovation, more is needed from private investors. Interest in investing venture capital is growing, particularly following a favourable change in capital-gains tax in 1999. But much more is needed. A survey by the Australian Bureau of Statistics published in March revealed that, as of 30 June 2000, only 5% of the country's A\$4.9 billion venture-capital investment was in biotechnology. New biotech start-ups often lack business or marketing expertise, which is a deterrent to investors. To draw deeply from the well of global monetary funds, biotech companies will need to demonstrate that they have the strategic planning

in place to take a product from invention through to the international market. And many research-industry consortia set up through government and state programmes would do well to adopt the management structure and operating practices of business enterprises such as the more industry-focused Cooperative Research Centres.

Commercializing research

The entrepreneurial spirit driving the US biotechnology industry is only just catching on in Australia. Universities, research organizations and individual scientists have often been reluctant to drive their discoveries through to commercial development, and have typically licensed their intellectual-property rights to another party. Nevertheless, a fundamental shift in attitude is occurring, and efforts to turn a scientific discovery into a business enterprise are gaining acceptability. Universities and research organizations must ensure that there are no barriers (subject to considerations of conflicts of interest) preventing researchers from taking an active role in commercializing their research.

Australia will reap higher returns if it keeps a stake in its biotechnology products until they are well advanced in their market development. This may in some cases mean importing commercial expertise and marketing skills. Such was the case of the biotech success story of the flu drug Relenza. Discovered in Australia, it was developed by a home-grown spin-off company, Biota, which strategically partnered with the pharmaceutical giant Glaxo Wellcome to take it through clinical trials, some of them in Australia. The example also provides a warning: over-hyped expectation has hampered its success on the stock market.

The country has major strengths that could make it a centre for international investment in biotechnology research and development. It has distinct cost advantages over other countries, particularly with the fall in the value of the Australian dollar in the past year. And its regulatory regime is aligned with practices elsewhere, especially those in the United States and Europe. For these reasons, and for its good public health-care system, pharmaceutical groups are increasingly seeing Australia as an excellent place for clinical trials. But the country's pharmaceutical industry could do much more to recognize and exploit opportunities there.

Similarly, the Australian biotech community must try far harder to sell its research initiatives overseas. With an Australian population of barely 20 million, a biomedical product that required substantial investment to reach the market-place will clearly need to penetrate the international market. So although the Australian biotech sector should be protective of its intellectual property, it must also have the know-how to become globally competitive.

But it is up to the scientific community and entrepreneurs, while making the most of government investments, to lobby to improve the universities, which have suffered considerably in recent years. If Australian biotechnology is to prosper in the long term, stronger support for universities, and perhaps also some high-profile success in biotechnology itself, will be necessary to enhance the currently low appeal of science to Australian students entering university. ■





Hot prospect
Patent loss opens way
for cheaper PCR
enzymes
p622



Worldly wise
US physicists
encouraged to think
internationally
p623



Museum trips
Smithsonian
reforms face
fresh controversy
p624



Rocky road
Samples from Mars
set for spell in
quarantine
p625

Researchers unnerved by echoes of the past in Russian directive

**Bryon MacWilliams, Moscow
and David Adam, London**

Foreign links with Russian science are secure — at least for the time being — despite last week's emergence of a directive from the Russian Academy of Sciences requiring all researchers' contacts with foreigners to be reported, in detail, to the academy.

Sergei Kovalyov, a member of parliament



Sergei Kovalyov:
concerned over
the directive.

who made the contents of the directive public in a radio interview on 31 May, claimed that its existence reflected a revival of the state security apparatus of the Soviet era.

But researchers inside and outside Russia say the directive need not dampen collaboration between Russian and foreign scientists. Some of them cau-

tion, however, that its existence may reflect a trend towards less open scientific exchange under Russian President Vladimir Putin.

The directive instructs the academy's departments and its 440 research institutes to tighten controls on the foreign contacts of their 53,000 scientists — ostensibly to curtail the loss of state secrets through espionage.

Kovalyov, a noted human rights activist, says that the directive reflects the security outlook of Putin, a former KGB officer. "He who assumes power gets to choose the music. Members of the KGB have assumed power. They are doing that which they were taught to do," he says. After it was leaked, some observers, including the financier George Soros, who has supported extensive research efforts in Russia since the end of the Cold War, expressed alarm at its likely implications.

But Igor Milovidov, an assistant secretary of the academy, describes the directive as "an ordinary little reminder". And Eugene Sverdlov, director of the Institute of Molecular Genetics in Moscow, says: "I cannot see any threat to scientific freedom or the exchange of ideas in this. We received an instruction to indicate who we met during trips abroad from the academy maybe two years ago. Time has passed and nothing has happened. People still

go abroad as freely as they did before."

Zhores Medvedev, an exiled Russian geneticist living in London, suggests the reason for the directive is bureaucratic. "I don't think this is a political move," he says. "It was necessary to establish certain kinds of rules because nobody was telling the academy what kind of grants they received from abroad."

But other Russian scientists suspect that the directive does represent an effort by the Russian Academy of Sciences to reassert some control over its staff. Maxim Frank-Kamenetskii, a biophysicist at Boston University, says: "I think this is an attempt by the leadership [of the Academy of Sciences] to regain some of the control it lost over recent years."

The directive comes at a time when several Russian civilians have been arrested on espionage charges. A researcher specializing in arms control is currently on trial for allegedly spying for unnamed NATO member countries, and a physicist is awaiting trial on suspicion of spying for China. "Trials linked to the so-called excessive freedom of scientific research are not so uncommon for us now," Kovalyov told the radio station.

The greatest impact of any clamp-down on foreign links will be on those academy institutes that have thrived since the 1991



Control centre: the Russian Academy of Sciences is seeking tighter regulation of foreign contacts.

collapse of the Soviet Union. These have struck deals with foreign laboratories and businesses, and funds from abroad make up significant portions of their budgets.

Watchdog fuels doubts over laser

Rex Dalton, San Diego

A Congressional watchdog agency has strongly criticized the US Department of Energy (DOE) for the way it has managed the National Ignition Facility (NIF), the troubled laser fusion project being built at the Lawrence Livermore National Laboratory in California.

The General Accounting Office (GAO) issued a report on 1 June saying that NIF has not yet had a fully independent review of its costs and prospects for technical success. The agency adds that the DOE has not hired the necessary staff to manage the project, and that significant technical questions remain about NIF's likely contribution to

the nuclear weapons 'stockpile stewardship' project, of which it is part.

NIF, a stadium-sized laser device that is supposed to ignite fusion in a tiny pellet of tritium-deuterium fuel, is now scheduled to be completed in 2008, six years later than planned, at a total cost of \$4.2 billion, or about \$1.4 billion over its initial budget.

James Anderson, the DOE's project officer for NIF, says the GAO report is generally fair, but takes issue with the need for an independent review. "We are satisfied NIF has been adequately reviewed," he says. He called the hiring of more scientific management staff for the project "a work in progress".

Patent ruling could cut PCR enzyme prices

Rex Dalton, San Diego

The European Patent Office (EPO) last week revoked a patent for an important thermally stable enzyme used in the polymerase chain reaction (PCR) process for DNA amplification. The decision brings with it the prospect that the price of the enzyme will fall.

Hoffmann-La Roche's patent for naturally occurring *Taq* DNA polymerase — obtained from the bacterium *Thermus aquaticus*, which lives in hot springs — is invalid because it is not a novel invention and is based on previously published discoveries, EPO officials say. Roche says it will appeal against the decision next year once the formal written opinion is issued by the EPO in Munich.

Melinda Griffith, general counsel for Roche Molecular Systems in Pleasanton, California, downplayed what she called "an intermediate-level" decision. She said the panel of patent examiners responsible had made "many errors of law" during the three-day hearing, which ended on 30 May.

This is the second major Roche patent on PCR to be revoked. In 1999, a federal judge in San Francisco ruled that Roche's US patent for native *Taq* was invalid (see *Nature* 402, 709; 1999) because it had been obtained with "intent to mislead". Switzerland-based



Getting warmer: Tom Brock of the University of Wisconsin, who discovered *Thermus aquaticus*.

Roche is appealing against that decision, and a court ruling is expected next year.

In both the EPO and San Francisco cases, the Roche patents were disputed by Promega, a Wisconsin-based company that markets native *Taq* and other reagents internationally. Three other companies —

Bioline, New England Biolabs and Becton Dickinson — supported Promega in the EPO action. The companies argued that the EPO patent for native *Taq*, issued in 1997, should never have been allowed and was granted only after years of pressure by Roche and its subsidiaries.

Roche still holds an EPO patent for a recombinant version of *Taq*, which is created by inserting the gene for *Taq* into another species of bacterium. Recombinant *Taq* now accounts for a large proportion of the market. But that patent is being challenged by the London-based company Bioline.

Bioline declined to comment on the case. But Griffith claims Roche's EPO patent on recombinant *Taq* is valid and enforceable.

In a separate action, Roche is suing Promega in the German federal court for selling *Taq* in Europe, alleging that the company is infringing its patents. It remains unclear how the EPO ruling will affect this action.

Promega lawyer Brenda Furlow claims that more companies are now likely to start selling native *Taq*, creating a more competitive market that will lower the price of the widely used reagent. Native *Taq* produced by the companies fighting Roche costs 20–30% less than Roche charges, she notes. ■

French unions walk out of talks over 35-hour week

Sally Goodman, Paris

Unions representing French researchers walked out of a meeting at the Ministry of Research last week in protest at the government's plans to introduce a 35-hour week. The dispute centres on the government's refusal to create new positions to compensate for the working time lost.

"Any measure to reduce working hours should be accompanied by new posts, otherwise our research potential will

decrease," warns Jacques Fossey, secretary general of the scientific researchers' trade union SNCS-FCU. The unions will decide what action to take after a further meeting with government officials on 26 June. Other public-sector unions have called strikes.

Fossey estimates that most researchers work between 45 and 50 hours a week on 39-hour contracts. "For researchers, working a 35-hour week has no real meaning," he says.

The French public research agencies, such as the CNRS (the national agency for basic research), are being asked to implement a new law by January 2002 which stipulates that all civil servants should work a 35-hour week, or a maximum of 1,600 hours per year, with extra hours being compensated by additional days off. The legislation will also cover university staff.

"We think the research agencies are being offered a good deal," says a ministry spokeswoman. "Staff will be able to save up their extra days off over five years and take extended holidays, and there will also be financial compensation." She adds that new posts are budgeted for in 2002, although these are not directly linked to the new law.

Abderrahmane Tadjeddine, director of

the LURE synchrotron laboratory in Orsay, near Paris, says that the main problem is implementing a system that will keep machines operating round the clock.

In the private sector, the 35-hour week has established itself firmly as part of French culture since its introduction over a year ago. Many observers say that it has fulfilled the government's stated objective of creating new jobs. A recent government survey found that 59% of those affected say their lives have improved.

But applying the law to civil servants, of which there are nearly 5 million in France, is proving problematic. The government is standing firm on its decision that no new jobs will be created in the public sector to compensate for the reduction in hours.

The situation is complicated by the fact that research agencies such as the Atomic Energy Commission (CEA) and the French national space agency (CNES) have industrial rather than civil-service status and so have already introduced the shorter working hours. The CEA has announced that 500 new posts will be created over three years, but this has been achieved only after difficult negotiations and a salary freeze. ■



Review urges US physicists to think globally

Irwin Goodwin, Washington

In a major review of the entire field of physics, a panel of the US National Academy of Sciences has called on physicists to promote international collaboration, and to make more of the importance of applications of their work to the life sciences.

The panel's report, *Physics in a New Era*, urges Congress and federal agencies to develop new protocols for international cooperation. "A key issue is the importance of long-term, stable funding for large-scale international projects," it says.

The issue is critical to US physicists, following the collapse of the Superconducting Super Collider project in Waxahachie, Texas, in 1993, and the 1999 abandonment of US involvement in the International Thermonuclear Energy Reactor, a global collaboration in fusion research.

The call for collaboration is one of nine main recommendations in the report, to be released later this month. The panel also recommends that funding for basic physics be restored, as a share of the economy, to the level it enjoyed 20 years ago, requiring current spending to rise by about one-third.

It argues that "along with the high-tech economy, the biological and medical sciences have benefited enormously from this research. Such physics-invented technologies as X-ray crystallography, magnetic resonance, fibre optics, electron microscopy, mass spectroscopy and radioactive tracers are at the heart of the rapid pace of discovery in almost every biomedical laboratory."

The panel is mildly critical of university physics departments, urging them to review and revise their curricula, "and to reverse... the long-term decline in the numbers of undergraduate and graduate students studying physics".

It calls on science agencies that fund physics — the Department of Energy, the National Science Foundation, NASA and the Department of Defense — to support the work of more individual investigators. It says that funding for small groups and single investigators is "dangerously inadequate".

It also warns that security problems at the Department of Energy's nuclear-weapons laboratories have produced an adversarial climate between scientists and managers. "The morale of scientists is low, and recruitment and retention difficulties threaten the viability of the laboratories," the panel says.

The report is an overview of a ten-year study by the academy that has spawned six smaller reports on subdisciplines of physics. Unlike its influential decadal reviews of astronomy, for example, the academy's physics review does not set priorities for funding of specific facilities or projects.

"Ours is an order of magnitude different



Local enterprise: new mechanisms are needed to enable international physics projects to match the success of US-only ones, such as Brookhaven's Relativistic Heavy Ion Collider (above).

from the astronomy reports," explains Yale University's Tom Appelquist, chairman of the 14-member panel. "We speak about the opportunities for discovery that cut across the expanding diversity of physics. Our priorities are not a wish list."

"It's devilishly difficult and divisive to set priorities in a report that encompasses all of physics," says James Langer of the University of California, Santa Barbara, last year's president of the American Physical Society.

The earlier subpanel reports did endorse particular projects: particle physicists, for

example, recommended capitalizing on the potential of Fermilab's main injector. Nuclear physicists supported Brookhaven's Relativistic Heavy Ion Collider and promoted an as-yet-unbuilt rare-isotope accelerator, and condensed-matter physicists championed the Spallation Neutron Source in Oak Ridge, Tennessee.

The overall report calls on US physicists to "develop a broadly shared vision" of experimental research facilities and to "communicate this vision clearly and persuasively" to politicians and the public. ■

MIT to set up Media Lab in India

K. S. Jayaraman, New Delhi

The Indian government has approved a ten-year, \$1.1-billion joint project with the Massachusetts Institute of Technology (MIT) to set up Media Lab Asia, modelled on the highly successful Media Lab in Cambridge, Massachusetts.

The Indian facility will be MIT's second Media Lab outside the United States, following the establishment last year of Media Lab Europe in Dublin, Ireland.

Announcing the agreement on 31 May, Indian information-technology minister Pramod Mahajan said that MIT had chosen India after evaluating offers from several other Asian countries, including China and Singapore. A formal agreement between his ministry and MIT is expected to be signed this month.

Mahajan said the project would initially have a one-year "exploratory phase", for which the Indian government has committed \$14 million. Whether it

continues for the next nine years will then depend on an end-of-year review.

If it continues, the government will provide 20% of the funding, with the rest expected to come from industry. Mahajan said the goal is to make the new laboratory self-sustainable, with income from the commercialization of its research findings.

Although conceptually similar to its cousins in the United States and Europe, Media Lab Asia will have a more decentralized structure, with offices in Mumbai, New Delhi and Bangalore. Mahajan says it will focus on technologies and products that are of specific value to Asia.

In particular, it will develop ideas for future global and domestic markets, and seek affordable solutions to problems of health education and rural development. One of its first tasks will be to identify academic partners to work with entrepreneurs on projects to benefit rural areas. ■

Rivals clash over plans to take science to London public

David Adam, London

Researchers planning an £8 million (US\$11 million) science communication centre at London's Science Museum have been wrong-footed by a separate proposal to open a similar centre, more than a year earlier and only a few miles away.

Recent high-profile controversies over such issues as the risks posed by bovine spongiform encephalopathy and genetically modified crops have made science communication a big issue in Britain. Both initiatives aim to engage scientists in a dialogue with the public.

Later this month, officials at the Science Museum will announce plans to build a new venue for people to meet scientists and discuss controversial issues such as cloning and animal experiments. The building will feature lecture and discussion rooms, as well as informal social facilities. The British Association for the Advancement of Science, which promotes science communication and education in the United Kingdom, will move into the building when it opens in 2003.

Meanwhile, the Royal Institution, an independent organization also involved in science communication, is planning to open a 'science media centre' early next year. Susan Greenfield, director of the Royal Institution, says the centre will help scientists discuss their work and its risks with the public by acting as a conduit to journalists. She plans to put a 'science spokesperson' in charge of the centre with the authority to speak to the press on behalf of media-shy colleagues.

But some scientists behind the museum's project say Greenfield failed to consult them about the Royal Institution's plans — and that, as a result, both are trying to do much the same thing.

"It is unfortunate that there was not sufficient prior consultation about the Royal Institution's media centre in light of the already well-developed plans for the new Science Museum building," says Colin Blakemore, chair-elect of the British Association.

Others are also critical of the Royal Institution's plans. Aspects of the proposals, such as a one-stop science spokesperson to comment on everything from AIDS to astronomy, are "unwelcome and unworkable," says one official of a leading scientific society.

Greenfield admits that she could have consulted more widely, but insists there is room for both projects. ■

Museum director quits over Smithsonian restructuring



Changes due: plans include separating the management of research from that of exhibitions.

Corie Lok, Washington

Plans to streamline research at the Smithsonian Institution, the world's largest museum complex, remained steeped in controversy last week with the resignation of the director of its National Museum of Natural History.

Robert Fri quit his post at the museum, which contains the institution's largest research operation, apparently in protest at the proposed restructuring.

In an effort to defuse the mounting tension between management and scientists over its plans, the Smithsonian's administration will this week name a 15-member committee to advise it on the details of the restructuring.

The panel will be chaired by Jeremy Sabloff, director of the University of Pennsylvania Museum of Archaeology and Anthropology in Philadelphia. It will feature members drawn from both inside and outside the Smithsonian.

The institution's governing Board of Regents voted to set up the advisory panel last month (see *Nature* 411, 119; 2001), after Smithsonian secretary Larry Small and science undersecretary Dennis O'Connor were criticized for pushing their restructuring plan through without consulting scientists.

According to Smithsonian officials, that plan has now been put on hold, pending the advisory panel's recommendations. But the institution's scientists remain sceptical of the

panel's influence and believe that parts of the plan, including the separation of the management of research from that of exhibitions, are almost certain to be adopted.

In his resignation statement Fri said: "It seems likely that the research and collections responsibilities of the museum will report directly to the undersecretary for science." Restructuring the museum "will require the leadership of a management team committed to pursuing its success over the long haul", he continued. "I do not feel that I can make that commitment enthusiastically." Fri will stay on as director until the end of September.

According to Brian Huber, a palaeobiologist at the National Museum of Natural History and chair of the Smithsonian's academic senate, the separation of research and exhibits would jeopardize the Smithsonian's mission of increasing and disseminating knowledge. Exhibits require input and interaction with scientists, he says.

Such a separation was discussed at a meeting of the Board of Regents last month and is seen as "a good idea", according to board member Manuel Ibanez, a retired biologist and former president of Texas A&M University in Kingsville. Bringing research under one director could increase efficiency without preventing interaction between scientists and exhibit staff, he says.

But no final decisions were made on the proposal and the board remains open to the advisory panel's recommendations, according to Hanna Gray, another board member, a historian and former president of the University of Chicago.

Huber says that an ad hoc group of scientists, of which he is part, is putting together its own restructuring proposal and will submit its plan to the advisory panel later this summer.

Sources at the Smithsonian say that Rita Colwell, director of the National Science Foundation, was invited to co-chair the advisory panel with Sabloff but declined to do so. Colwell would not comment.

Meanwhile, the institution's restructuring plans, which must be approved by the Congress, came under attack on another front, with two Maryland senators writing to Small protesting over the proposed closure of the Smithsonian Center for Materials Research and Education in Suitland, Maryland. Small backed down last month from closing the Conservation and Research Center in Front Royal, Virginia, in the face of opposition from researchers and Virginia congressmen. ■

Mars rock samples condemned to quarantine

Tony Reichhardt, Washington

As if collecting rocks from the surface of Mars and returning them to Earth weren't challenge enough, a panel of scientists is recommending that NASA design and build a unique quarantine facility to ensure that any organisms in the martian samples don't contaminate the terrestrial environment.

Quarantine of martian rocks has long been considered a requirement for any mission to collect samples. But last week's report by the National Academy of Sciences' Committee on Planetary and Lunar Exploration (COMPLEX) was the most detailed look yet at how such a facility should be designed.

The panel, chaired by John Wood of the Harvard-Smithsonian Center for Astrophysics, says that NASA should start working on the problem as soon as possible. The space agency currently plans to bring martian samples back to Earth by about 2014.

The chance of finding viable organisms in a martian rock is extremely remote, but scientists have nonetheless advised utmost caution in handling returned samples.

The report recommends that the Mars quarantine lab apply the most stringent level of control, Biosafety Level 4 (BSL-4), which is typically used for germ-warfare agents or pathogens such as the Ebola virus.

The simplest solution would be to build the new facility next to an existing BSL-4 lab, such as those at the Centers for Disease Control and Prevention in Atlanta and at the US Army Medical Research Institute of Infectious Diseases at Fort Detrick, Maryland. Another BSL-4 facility is being built at the

University of Texas at Galveston, near NASA's Johnson Space Center, which houses the Moon rocks from the Apollo missions.

The new facility, however, will have to function both to keep contamination out and to contain pathogens. The usual way to accomplish one of these is to use rooms with different atmospheric pressures, so that leakage is forced in the desired direction. But no one has built a lab that does both.

The COMPLEX panel did not discuss engineering details, but Wood is confident that "it can be done". The price tag is largely guesswork, but based on the \$8 million it cost NASA to build a laboratory for the Moon rocks in the 1960s, and adjusting for inflation, Wood estimates that the facility could be built for around \$32 million. NASA plans to spend \$427 million this year alone on its Mars exploration programme.

The panel recommends that the facility be kept as simple as possible, and that inves-

tigation of the samples take place elsewhere. So the other challenge facing the lab's designers is how to sterilize samples for release without destroying any biological evidence.

The recommended method for sterilizing samples is gamma-irradiation and dry heat. But heating the samples would remove water and destroy organic compounds. The answer may be to extract compounds first, sterilize them, and send them out for study.

But the panel admits that more research is needed. "Unfortunately the recommended techniques of sterilization damage organic compounds to some extent," it says. "It is important that studies be carried out to enlarge our knowledge in this area."

In the unlikely event that the rocks contain unambiguous signs of life, they would stay in the quarantine facility until a research strategy could be formulated. But that possibility is so remote that COMPLEX did not provide a detailed plan for it.



Beware of the martians: rock samples could be given the same biohazard status as germ weapons.

JPL/NASA, MALIN SPACE SCIENCE SYSTEMS

Aurora project set to pave way for human space flight

Alison Abbott, Munich

The European Space Agency (ESA) is planning a programme that will enable it to maintain an interest in human space flight, once the construction of the International Space Station is completed in around 2005.

The programme, called Aurora, won support at a meeting of representatives of ESA member states last week. It would probably involve an initial series of robotic missions — mainly to asteroids, Mars and the Moon — followed by human missions.

An important branch of Aurora would develop technologies for human space travel. Research would address soft spacecraft landings, and solar-electric and nuclear propulsion systems to allow astronauts to travel faster — thereby reducing their exposure to cosmic radiation. Physical and psychological health research programmes for astronauts would also be included.

But another part of Aurora would support science, including exobiology — the study of the origin, evolution and distribution of life in the Universe.

There is no doubt about the scientific community's enthusiasm for the Aurora concept, ESA officials say. The agency put out a four-week call for ideas for the proposal in February and received nearly 300 submissions — "ten times the number we normally attract," says Franco Ongaro of ESA's Advanced Concepts and Studies Office.

"This is just the right time to ride the wave of interest in exobiology," says Paul Murdin of the British National Space Centre, the United Kingdom's delegate to the Aurora committee. Murdin says the discipline is attracting a whole new community of researchers.

ESA will ask its governing council, which meets this month, to support a 40-million-

euro (US\$34 million) three-year preparatory phase for Aurora, starting next year. ESA member governments will probably scale this down, officials say. Each member state will be able to choose whether or not to participate in the programme, although it seems likely that most of the agency's 15 member nations will sign up to it.

Aurora would be managed separately from ESA's existing programmes for science and human space flight. There is concern that it could threaten funding for these programmes, which is mandatory for all ESA members. But Murdin says that it would complement existing programmes and enable ESA to develop the technologies needed for human space flight.

Supporters of Aurora envisage that it could eventually match ESA's existing science programme in size, with an annual budget of 200 million–300 million euros.

Ecologists call for greater scrutiny of GM organisms

Washington The environmental effects of genetically modified organisms (GMOs) require greater scrutiny than those of organisms produced by traditional breeding practices, the Ecological Society of America (ESA) said last week.

The ESA, which represents 8,000 US ecologists, says the difficulty of studying and predicting the long-term ecological effects of GMOs warrants caution over their use. Risks include the creation of new or harder pests, harm to non-target organisms, and loss or changes to biodiversity. "We really need more peer-reviewed research on the potential environmental effects of GMOs," says Diana Wall, an ecologist at Colorado State University and chair of the ESA committee that drafted the statement.

Genetically modified crops such as corn and soybeans are widely grown in the United States for human consumption. The GMO debate has often pitted pro-biotechnology scientists against environmental groups, but the ESA statement indicates that the scientific consensus in favour of the technology is far from complete.

Hopes dashed as foot-and-mouth rumbles on

London Fresh outbreaks of foot-and-mouth disease in Britain have ended hopes that the epidemic is almost over.

After hearing from three independent teams of epidemiologists, the government's chief scientific adviser, David King, announced last month that the epidemic would be effectively over by the time this week's general election took place. But further outbreaks in Yorkshire have now led King to state that the disease is likely to continue until August.

Epidemiologists working for the Ministry of Agriculture, Fisheries and Food warn that

it is difficult to predict how the tail of the epidemic will progress and that new cases are likely to be recorded for several months.

Cornell finally kills off nuclear reactor

New York Cornell University's nuclear reactor is to be decommissioned after nearly 40 years. A university panel has concluded that it was no longer academically and economically feasible to continue using it.

The 500-kilowatt reactor at Cornell's Ward Center for Nuclear Sciences, in Ithaca, New York, is the only such facility in the state. It will be dismantled over the next few years in a process expected to cost millions of dollars. A gamma-ray facility at the centre will be maintained. Cornell disbanded its nuclear science and engineering programme in 1995 because of decreasing interest in nuclear power, but the reactor has continued to be used for research and training purposes.

Framework programme aims to boost big science

Brussels The European Commission last week released details of how it intends to allocate funds for the next Europe-wide research programme, Framework 6, which is set to begin in 2003.

As expected, the 17.5 billion euros (US\$15 billion) for the four-year programme will be concentrated on large-scale projects in seven priority areas of research. Top of the list are information technology with 3.6 billion euros, and genomics and biotechnology, set to receive 2 billion euros.

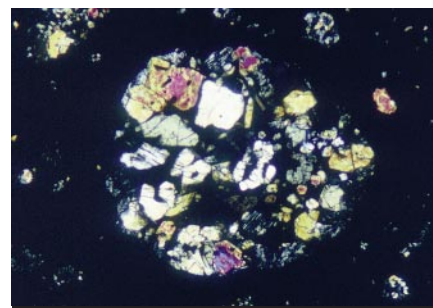
The commission hopes the new scheme will help to integrate research capacities across Europe, and pave the way for a pan-European research area (see *Nature* 410, 4; 2001). The Council of Ministers will discuss the proposal later this month.

Australians dismayed by share of budget

Canberra Australian scientists reacted with disappointment to last month's national budget, claiming that the money allocated to science fails to make up for previous cuts.

The A\$160-million (US\$80-million) allocation is the first part of a five-year, A\$2.9-billion spending plan for science announced this February (see *Nature* 409, 655; 2001). Researchers had hoped for more now, but Treasurer Peter Costello chose to divert money into other programmes, including schemes to help the elderly.

"The chance to kick-start universities has been ignored," says Michael Barber, policy secretary of the Australian Academy of Science. Government officials say that further increases in science spending would require taxes to be raised.



Meteor mystery: how did the fragments of Cold Bokkeveld meteorite end up in private hands?

Natural History Museum in row over meteorite sale

London Mineralogists at London's Natural History Museum say that they were not involved in the sale of a sample of a meteorite that was found in South Africa and that, some say, is the rightful property of the South African government.

The South African *Sunday Times* reported on 3 June that it had been offered fragments of the Cold Bokkeveld meteorite, which fell to Earth in Cape Province in 1838, by Mike Farmer, a dealer based in Tuscon, Arizona. The South African government is currently attempting to crack down on illegal sales of items belonging to the country.

A Natural History Museum spokesperson says the fragments were donated to the museum in 1839 by Sir John Herschel, who was chief astronomer at the Cape Observatory. Museums regularly exchange samples with other museums and dealers to improve their collections, but the spokesperson was unable to confirm immediately whether the fragment in question had been exchanged with Farmer or with another dealer.

Gates gift boosts hunt for meningitis vaccine

Washington The Bill & Melinda Gates Foundation has donated \$70 million to help develop a meningitis vaccine for sub-Saharan Africa.

Existing vaccines used in the area are not effective in children under the age of two, the group most at risk of infection, and only protect older children for a limited period. A longer-lasting vaccine is used in Britain, but it does not protect against one of the two meningitis serotypes found in sub-Saharan Africa. Meningitis infects up to 250,000 people a year in the 'African meningitis belt', an area stretching from Ethiopia in the east to Gambia in the west.

The funds will be administered by the World Health Organization and the Seattle-based Program for Appropriate Technology in Health, a non-profit public health organization.



Spread betting: still no end in sight as foot-and-mouth disease reaches the Yorkshire Dales.

Tabletop astrophysics

Can a bowlful of cold atoms help physicists simulate some of the most extreme conditions in the Universe? Philip Ball goes on the trail of the laboratory-scale black hole.

A new chapter has been added to the physics cook book. The recipes may be simple — take a gas of atoms, cool and stir — but the dishes that should result could hardly be more surprising: miniature black holes, white dwarfs and supernovae.

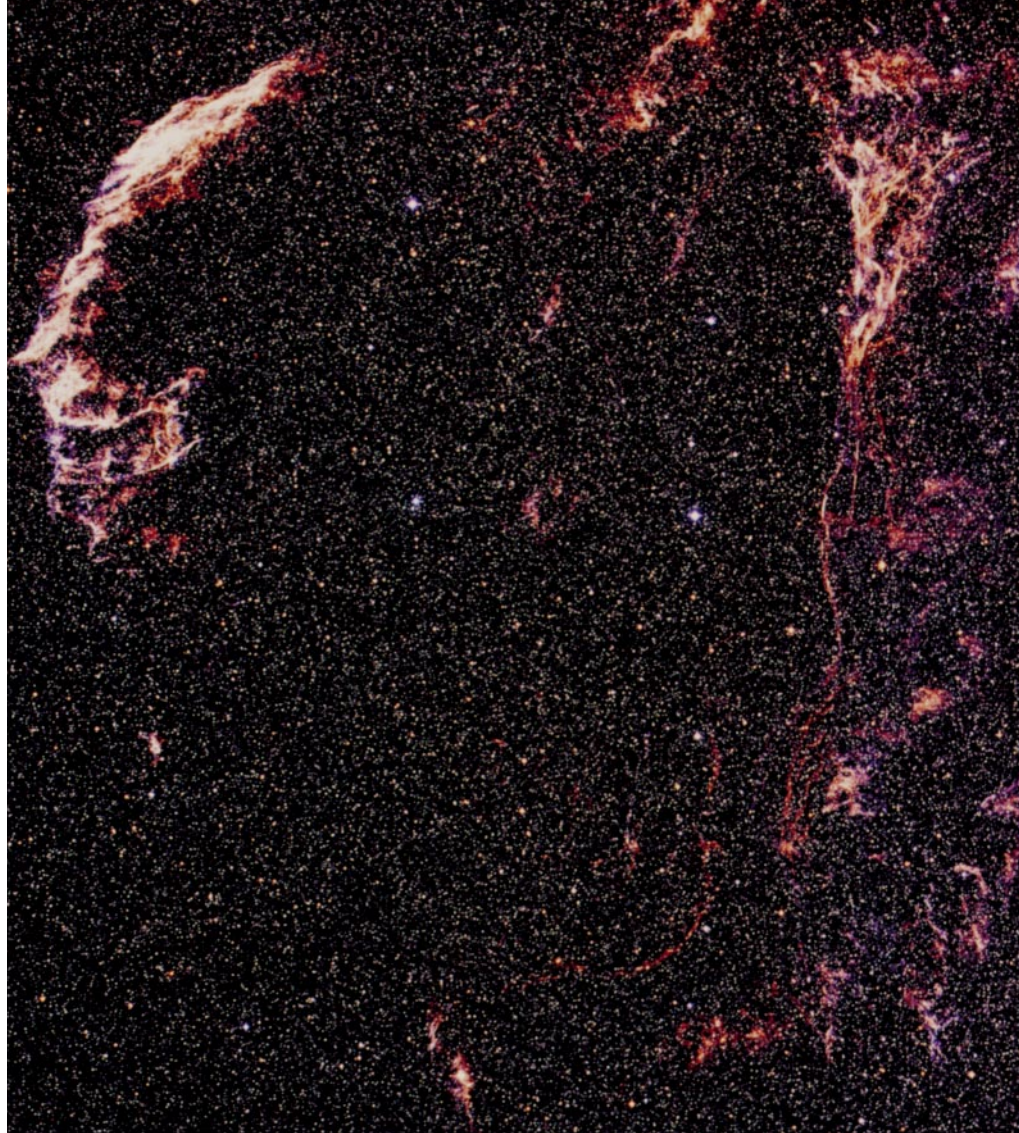
Key to this culinary magic is a strange state of matter that only exists in laboratories, known as a Bose–Einstein condensate (BEC). In 1995, when this state was first seen in a gas of cold atoms¹ — over 70 years after BECs were predicted to exist² — it caused a storm in the world of fundamental physics. But outside this arena, observers could be forgiven for asking what these strange condensates could be used for.

In recent months, an answer may have been found. Physicists have shown that BECs can mimic some of the properties of collapsing and exploding stars. So in much the same way that ripples on water can be used to study the properties of light waves, BECs might offer a way to bring extreme astrophysics into the lab.

All together now

BECs get their name from physicists Satyendra Nath Bose and Albert Einstein. In 1924, Einstein built on work by Bose to predict the existence of the condensates². They have their roots in quantum mechanics, the physical theory of the microscopic world. In quantum theory each particle exists in a particular ‘quantum state’, which defines, for example, the particle’s energy.

But when it comes to sharing quantum states, some particles are more friendly than others. All particles are divided into two types: fermions and bosons. Fermions are distinctly antisocial: they refuse to gather together in the same quantum state. Bosons,



These models may be the best chance of recreating dying stars in the lab.

on the other hand, are remarkably gregarious, and it is this sociable nature that ultimately gives rise to a BEC.

Normally, thermal energy gives bosons a range of different energies and distributes them among different quantum states. Einstein showed that at low enough temperatures, there is no limit to the number of bosons that can exist in the same state. In other words, if you cool a group of bosons down sufficiently, they will all crowd into the lowest energy state. Quantum mechanics says that particles in the same state behave as a single particle, so the cooled crowd of bosons acts like a single collective particle — a BEC, sometimes described as a ‘superatom’.

A little over a decade later, Einstein’s predictions were invoked to explain the unusual behaviour of cold helium atoms. In 1938, Soviet physicist Pyotr Kapitza found

Out with a bang: the explosion of a supernova, the remnants of which are seen here, could be modelled in the lab using very cold atoms.

that liquid helium lost all its viscosity when cooled to 2 kelvin — if you set it moving in a flask it would swirl around for ever³. Such liquids were later christened ‘superfluids’. Just months later, German physicist Fritz London showed that Kapitza’s discovery could be explained by assuming that the helium atoms — which are bosons — had condensed into the same quantum state⁴. Unknowingly, Kapitza had created a BEC.

But BEC research did not really take off until 1995, when Carl Wieman and his colleagues at the University of Colorado at Boulder managed to create a BEC from gaseous atoms. They used new laser-cooling technology to coax a low-pressure gas of rubidium atoms to close to absolute zero and into a BEC¹. The rubidium atoms remain as a gas throughout the transition to a BEC, making the process easier to study than in liquid helium. Wieman’s experiment was soon duplicated in laboratories around the world as fundamental physicists rushed to explore this unusual form of matter.

BECs do not exist outside the laboratory — even interstellar space is too warm

CELESTIAL IMAGE/SPL

for them. So are they nothing more than interesting oddities? At the beginning of last year, Ulf Leonhardt at the University of St Andrews in Scotland and Paul Piwnicki of the Royal Institute of Technology in Stockholm, Sweden, argued otherwise. They pointed out that a BEC could, under the right conditions, start to behave like a black hole⁵.

Black holes are the remnants of old and massive stars. As a star ages, the pressure generated by the nuclear reactions at its centre diminishes, and it eventually collapses inwards under its own gravity. The collapse can result in spectacular explosions known as supernovae, which leave behind a dense core of matter. If the original star was five or six times larger than the Sun, this remaining matter becomes concentrated at a single point. This point, and the area around it, is called a black hole.

Once formed, black holes have a voracious appetite. They use gravitational attraction to suck in anything that strays closer than a certain critical distance — called the event horizon. Even passing beams of light are devoured if they get too close. Some astrophysicists compare this effect to a whirlpool sucking in fish. The speed of the water is greater closer to the centre of the whirlpool. If a fish gets too near, and enters an area where the water speed is faster than the fish can move, it will be sucked inwards.

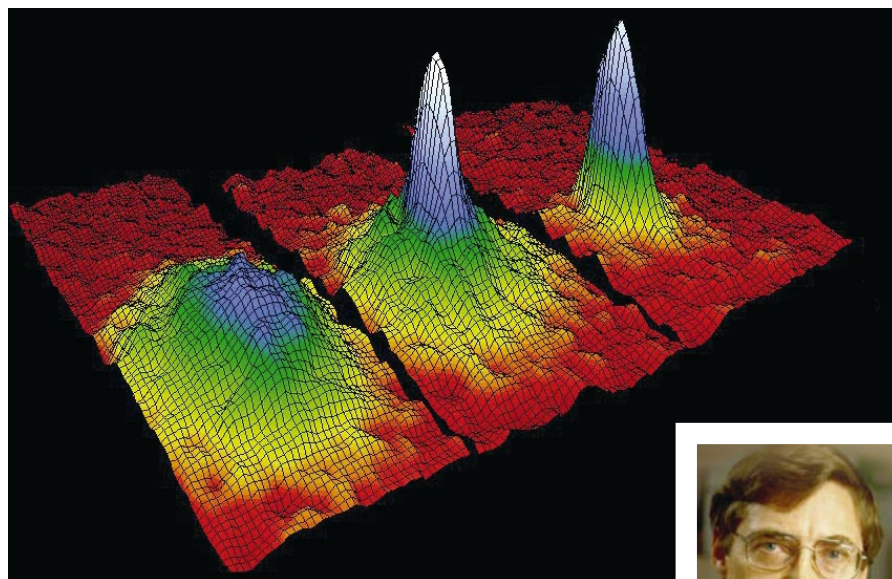
Spinning against the tide

In principle, black holes could be simulated by passing light close to a whirlpool of gas. If the centre of the whirlpool was moving faster than light, the light would be sucked inwards like the unfortunate fish. But creating a whirlpool that moves faster than the speed of light is a tall order. Unless, that is, light can be slowed down. And, over the past two years, physicists have managed to do just that using BECs.

In 1999, Lene Hau and her colleagues at the Rowland Institute for Science in Cambridge, Massachusetts, manipulated a cold vapour of sodium ions to form a BEC that slowed light to a trifling 17 metres per second⁶. Hau's group and, independently, Ronald Walsworth and his colleagues at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, have since brought light to a complete halt^{7,8}.

Leonhardt and Piwnicki suggested that if light was moving slowly enough, it could become trapped in a whirlpool within a BEC⁵. At a certain distance from the centre of the vortex, analogous to a black hole's event horizon, the light would circulate inexorably inwards. The swirling gas would become an optical black hole.

So far, no group has attempted to do the experiment Leonhardt and Piwnicki suggest. Ordinary BECs are difficult enough to create and maintain. The challenge of engineering one with a whirlpool in the middle is beyond



Get together: the peak appearing in these images shows the formation of a Bose-Einstein condensate in a gas as the atoms slow down and group together. It was first observed by Carl Wieman (right) and his team.



present technology. But there might be an easier way to mimic black holes. According to a team led by Ignacio Cirac and Peter Zoller of the University of Innsbruck in Austria, sound, instead of light, could be used to model black holes in a BEC⁹.

This 'sonic black hole' is similar to Leonhardt and Piwnicki's plan — a BEC is stirred to create a vortex that is moving faster than the speed of sound, trapping sound waves in the system. Because the BEC's bosons behave as a single particle, the swirling condensate does not contain any areas of turbulence that could obscure the effects physicists are looking for.

But stirring a BEC into a whirlpool is not easy. If the BEC is continually spiralling inwards, there must be some outlet that, like a plughole, removes atoms at the centre. This is hard to achieve, so Cirac and Zoller propose a different arrangement. The BEC is

confined to a thin ring, around which it flows continuously.

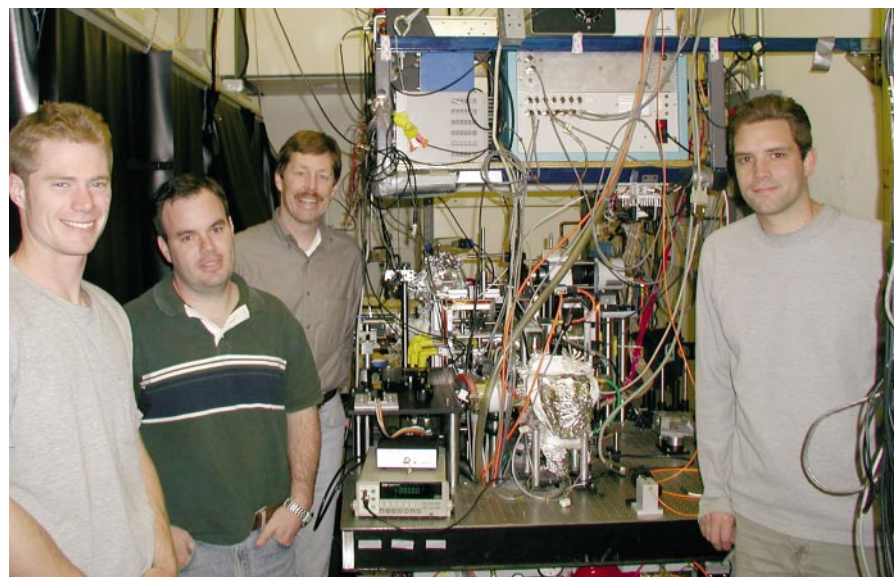
But one segment of the ring is thinner than the rest. The BEC is forced to speed up as it passes through the constriction into the thin part, accelerating it to supersonic speeds. The constriction is the sonic black hole's event horizon — sound passing this point cannot escape back out of the thinner area.

Zoller and his colleagues say that creating such a situation in a BEC is a delicate matter, but not impossible. "We believe that it may be carried out with present or near-future technology," says team member Luis Garay of the Institute of Mathematics and Fundamental Physics, an interdisciplinary research centre in Madrid run by Spain's national research council (CSIC).

BECs could also be used to study supernovae themselves, rather than the black holes



Into the abyss: a whirlpool in water provides a good analogy for the effect black holes have on matter.



Star-gazers: Randall Hulet (second right) and his team have mimicked white dwarfs in the lab.

▶ that they sometimes leave behind. Shortly after Wieman's team produced the first gaseous BEC, Randall Hulet and his colleagues at Rice University in Houston observed a dramatic collapse in the BEC system they were working on¹⁰. Two sets of forces can act on the atoms in a BEC. Although they act as a single particle, the individual atoms in a BEC continue to move slowly, and this motion results in a weak pressure that keeps the atoms apart. Depending on the type of atoms used, there can also be weak quantum-mechanical attractive forces acting between the atoms in a BEC. Wieman's group had used atoms that did not experience this attractive force, but Hulet's lithium atoms did.

Dance the Bosenova

Hulet's group discovered that if the number of atoms in their BEC became too great, the combined attractive forces caused the BEC to collapse. In a later experiment¹¹, the researchers watched the BEC go into a kind of oscillatory motion, as the collapsing atoms rebounded outwards, only for the attractive forces to pull them together again.

Wieman and his colleagues have since developed sophisticated methods for switching the atoms in a BEC between attractive and non-attractive states, allowing them to monitor individual collapses and rebounds¹². On close inspection, the collapsing BEC looks startlingly like a microscopic supernova.

The initial implosion is followed by an explosion in which the atoms are ejected as a hollow ball of atoms or in narrow jets, just as the collapsing star rebounds to form a characteristic expanding ball or streams of outflowing gas. In both a supernova and a 'Bosenova', as they are now known, a core of collapsed material is also left behind.

The big difference is in the amount of

energy released. A supernova liberates 75 orders of magnitude more energy than the exploding condensate, which musters only enough energy to raise its temperature by 200 billionths of a kelvin. But the BEC's collapse appears to be forceful enough to expel some of the atoms from the experimental system completely. As yet, the reason for the explosion is not clear.

Under pressure

Another fate of a dying star might also be open to investigation using very cold atoms. Eventually, our Sun will collapse to form a dense, blue-white body called a white dwarf. Such stars are spared further compression because of a balancing act between the electrons they contain and gravity. Electrons are fermions and cannot be squeezed into the same quantum state. This refusal to get too close to other electrons results in an outward pressure, known as Fermi pressure, that balances the inward pull of gravity. Unless the gravitational pull is strong enough to force the electrons to fuse with protons to make neutrons, a white dwarf will not collapse any further.

Earlier this year, Hulet and colleagues observed Fermi pressure in a gas containing two isotopes of lithium cooled to about 250 nanokelvin¹³. One isotope, lithium-6, is a fermion; the other, lithium-7, is a boson. Hulet found that clouds of the two isotopes, which were mixed but could be imaged separately, were different sizes and shapes as the lithium-6 atoms are affected by Fermi pressure whereas the lithium-7 atoms are not. The shape of the lithium-6 cloud, argues Hulet, arises from the same mechanism that stabilizes white dwarfs against further collapse.

The potential analogies for BECs in the lab are intriguing, but will astrophysicists learn anything from studying these systems?

A community is forming with lots of interdisciplinary connections.

Zoller thinks they can. One problem that they might help to shed light on, he says, is that of Hawking radiation. During the 1970s, Stephen Hawking of the University of Cambridge showed that black holes can become slightly warmer than the space surrounding them¹⁴. Black holes are no exception to the rule that heat flows from hot to cold, so they should, in theory, radiate energy — now known as Hawking radiation. This means that, given enough time, black holes should evaporate away altogether.

Hawking radiation poses a formidable challenge for physicists. It is too weak to be observed, but a theoretical description would have to combine quantum mechanics and gravity, and such a theory of quantum gravity does not yet exist. If an artificial black hole could be made, scientists could look for the presence of an analogue of Hawking radiation, providing useful evidence for the real thing.

Leonhardt points to other benefits, saying that the work on the astrophysical implications of BECs is bringing different scientific disciplines together. "A community is forming with lots of interdisciplinary connections between quantum optics, condensed-matter physics and astrophysics," he says.

If this community can make the analogies work, tabletop versions of astrophysical events could become reality. Whether or not these models will provide an insight into the real phenomena remains to be seen, but they may be the best chance researchers have of recreating the extreme conditions of dying stars in the lab. ■

Philip Ball is a consultant editor of *Nature*.

1. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. *Science* **269**, 198–201 (1995).
2. Einstein, A. *Sitzungsber. Kgl. Preuss. Akad. Wiss.* **1924**, 261 (1924).
3. Kapitza, P. *Nature* **141**, 74 (1938).
4. London, F. *Nature* **141**, 643 (1938).
5. Leonhardt, U. & Piwnicki, P. *Phys. Rev. Lett.* **84**, 822–825 (2000).
6. Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. *Nature* **397**, 594–598 (1999).
7. Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. *Nature* **409**, 490–493 (2001).
8. Phillips, D. F., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. *Phys. Rev. Lett.* **86**, 783–786 (2001).
9. Garay, L. J., Anglin, J. R., Cirac, J. I. & Zoller, P. *Phys. Rev. Lett.* **85**, 4643–4647 (2000).
10. Bradley, C. C., Sackett, C. A., Tollett, J. J. & Hulet, R. G. *Phys. Rev. Lett.* **75**, 1687–1690 (1995).
11. Sackett, C. A., Gerton, J. M., Welling, M. & Hulet, R. G. *Phys. Rev. Lett.* **82**, 876–879 (1999).
12. Cornish, S. L., Claussen, N. R., Roberts, J. L., Cornell, E. A. & Wieman, C. E. *Phys. Rev. Lett.* **85**, 1795–1798 (2000).
13. Truscott, A. G., Strecker, K. E., McAlexander, W. I., Partridge, G. B. & Hulet, R. G. *Science* **291**, 2570–2572 (2001).
14. Hawking, S. W. *Phys. Rev. D* **13**, 191–197 (1976).

Biology's name game

The confused nomenclature of genetics is blighting the field — some genes have multiple names whereas unrelated genes often share a common moniker. Helen Pearson examines attempts to bring order to the chaos.

Fruitflies have *armadillo*, mice have β -*catenin*. Both of these genes produce proteins that influence early embryonic development, and their DNA sequences are so similar that *Drosophila* and mouse geneticists agreed several years ago that they are basically one and the same. Yet the two names persist.

Fly geneticists, with their penchant for extravagant gene names, regard *armadillo* as a perfect label for a gene that, when defective, gives *Drosophila* embryos an armour-plated appearance. Mouse geneticists, meanwhile, see no reason to rename a gene that clearly belongs to the *catenin* family.

Genes with multiple aliases seem to be the rule, rather than the exception, whereas genes that have no functional relationship with each other can often bear the same names. As biologists strive to make sense of the growing wealth of genomic information, this messy nomenclature is becoming a bugbear. "You're

finding the closest relative of a gene is in a different organism," says Judith Blake, who works on the Mouse Genome Database at the Jackson Laboratory in Bar Harbor, Maine. Unfortunately, gene names are proving a hindrance, rather than a help, in making these connections.

Everyone wants to be able to associate related genes with one another in the databases. But how should this be done? Attempts to impose standard names across the board are meeting stiff resistance, and approaches that would give genes unique ID numbers seem unlikely to take off unless journals enforce the system. But a coalition of leading geneticists may have the answer. The Gene Ontology (GO) Consortium is sidestepping the naming issue by developing 'controlled vocabularies'. These will allow software to scan the genomic databases and link related genes to one another using terms that consistently describe their functions, regardless of what the genes are called.

Identity crisis

The extent of the nomenclature problem is illustrated by the work of a team led by Eivind Hovig at the Norwegian Radium Hospital in Oslo. Hovig and his colleagues are testing software designed to search for biological associations between genes based on their co-occurrence in the abstracts of published papers¹. By 1 May, their ongoing automated scan had examined more than

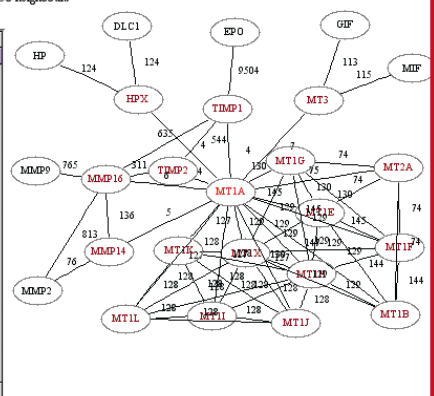
Genetic disorder: Eivind Hovig's studies have illustrated the naming confusion, including the multiple genes known as *MT1* (below).



MT1A
Found in 152 articles with 56 neighbours

Visit Neighbour

MT1B 147 times
MT1E 145 times
MT1F 145 times
MT1G 130 times
MT1H 129 times
MT1X 129 times
MT1J 127 times
MT1I 127 times
MT1K 127 times
MT1L 127 times
MT2A 75 times
MT3 7 times
MMP16 6 times
MMP14 5 times
TIMP2 4 times
HPX 4 times
TIMP1 4 times
GIF 3 times
MIF 3 times
MMP2 3 times
TP53 2 times
CRAT 2 times
IL2RA 2 times
CP 2 times
IL6 2 times



Biologists would rather share their toothbrush than share a gene name. **Michael Ashburner**

10 million records in the Medline literature database, identifying 22,008 distinct human genes. Of these, 10,352 had more than one name. One gene, officially designated as *SELL*, or *selectin L*, which controls cell adhesion during immune responses, had 15 aliases.

Even more confusing, 4,257 abbreviated names were used to refer to more than one gene. Top of the list was *MT1*, used to describe at least 11 members of a cluster of genes encoding small proteins that bind to metal ions.

But at least those genes are structurally related — which is more than can be said for the five unconnected genes that have at some point been referred to as *PAP*, involved in processes ranging from cell differentiation to inflammation of the pancreas.

Between organisms, the problem multiplies. Start reading any genetics review paper and you are likely to be juggling the names preferred for genes in different species within a few paragraphs. For example, the yeast homologue of the human gene *PMS1*, which codes for a DNA repair protein, is called *PMS2*; whereas yeast *PMS1* corresponds to human *PMS2*.

In an effort to address the problem, researchers working on the human and mouse genome projects have nomenclature



Going strong : Midori Harris coordinates the GO project from the European Bioinformatics Institute.

committees that are collaborating with one another and encouraging their respective communities to adhere to agreed names. But as researchers from a range of backgrounds jump on the genomics bandwagon, consensus is proving more difficult to obtain. "It's become a much bigger enterprise," says Sue Povey of University College London, who heads the Human Genome Organisation's Gene Nomenclature Committee.

Getting a consensus on gene names for the full range of important laboratory organisms could well be impossible. Used to working in relative isolation, researchers studying different species have grown attached to their respective gene-naming traditions. *Drosophila* geneticists, for instance, delight in using colourful names — such as *hedgehog*, which produces a signalling protein involved in a range of developmental processes, and *lost in space*, which guides the growth of neurons — and they have no intention of letting other geneticists spoil their fun. Two international nomenclature workshops, held in 1997 (ref. 2) and 1999 (ref. 3), concluded that, apart from between mammals, attempting to standardize gene names across species was pointless.

"Biologists would rather share their toothbrush than share a gene name," says Michael Ashburner, joint head of the European Bioinformatics Institute (EBI) at Hinxton near Cambridge, and one of GO's founders. "Gene nomenclature is beyond redemption."

There is also a realization that genes may have several functions as they are expressed at different times during development or are transcribed in different ways — so names based on known functions may turn out to

be misleading. "It'll be many years before we can agree on a set of names," says Mark Boguski, senior vice-president of research and development at Rosetta Inpharmatics, a company in Kirkland, Washington, that specializes in exploiting genomic information.

Genetics by numbers

One possible solution is to identify individual genes by unique ID numbers, and rely on database curators to provide links between related genes. The Mouse Genome Database already does this, assigning each gene an ID and listing all sequences deposited in the GenBank database that are related to it. Every entry in GenBank also has an accession number. But although journals such as *Nature*, *Nature Genetics* and *Science* demand that their authors list GenBank accession numbers in papers that describe a gene for the first time, they seem unlikely to move rapidly towards a system of enforcing the use of gene IDs or GenBank numbers at each mention of a gene. And without such a system, gene numbers seem unlikely to solve the nomenclature problem.

This is where GO comes in. Rather than trying to impose a standardized system for gene names or numbers, GO's members are developing agreed vocabularies to describe molecular functions, biological processes and cellular components^{4,5}. Using these terms, it becomes possible to link related genes irrespective of their muddled nomenclature.

For example, subunits of the transcription factor TFIIA, a protein that regulates gene expression, are encoded by genes with different names in *Drosophila* and yeast, but which share the same GO terms. Describing their molecular function as 'general RNA polymerase II transcription factor' means that the fruitfly genes *TfIIA-L*, *TfIIA-S* and *TfIIA-S-2* can be readily associated with their yeast functional equivalents, *TOA1* and *TOA2*.

GO was started by the curators of some of the main model organism genome databases — the *Drosophila* community's FlyBase, the *Saccharomyces* Genome Database, which is

used by yeast geneticists, and the Mouse Genome Database. It has since been joined by the curators of the database for the plant *Arabidopsis*, and WormBase, which serves researchers working on the nematode *Caenorhabditis elegans*.

Curators of the databases assign GO terms to individual genes and their products, and the information is fed back to a central GO database, maintained on servers at Stanford University in California. Ultimately, the consortium aims to produce a fully searchable database of terms that explain the function of genes in all organisms.

Deciding on appropriate terms has involved much discussion. For example, the polysaccharide chitin is a component of cell walls in yeast, but in *Drosophila* it is found in the insects' external cuticle. As a result, the GO term 'chitin metabolism', used to refer to genes involved in chitin production, now has two daughter terms, 'cuticle chitin metabolism' and 'cell wall chitin metabolism', to satisfy the requirements of fruitfly and yeast geneticists. "The vocabularies are dynamic — a work in progress," says Midori Harris, who in April was appointed as GO's first full-time coordinator, based at the EBI.

GO is also sufficiently flexible to accommodate synonyms where researchers commonly use two terms to refer to the same process — for example cell division and cytokinesis. Instances where the same name is used to refer to different processes by different research communities are harder to disentangle, says Harris. To a yeast geneticist, 'mating' means the fusion of two cells; whereas researchers working on mice would use it in the more conventional sense. GO's solution has been to adopt the term 'mating (*sensu Saccharomyces*)' — "doing it in the same way as budding yeast", explains Harris.

The GO system is rapidly gaining popularity, and GO terms featured in recent papers describing the *Drosophila*⁶ and human⁷ genomes, and a comprehensive library of mouse genes⁸.

The project has just received a major boost in the form of a three-year, \$5-million grant from the US National Human Genome Research Institute in Bethesda, Maryland, and has some enthusiastic converts. "We will stop purchasing products that don't use GO," says Ken Fasman, Boston-based global head of R&D Informatics with the drugs company AstraZeneca.

GO, it seems, is all systems go. ■

Helen Pearson works in *Nature's* science writing team.

1. Jensen, T.-K., Lægreid, A., Komorowski, J. & Hovig, E. *Nature Genet.* **28**, 21–28 (2001).
2. Blake, J. A. *et al.* *Genomics* **45**, 464–468 (1997).
3. White, J. A. *et al.* *Genomics* **62**, 320–323 (1999).
4. Ashburner, M. *et al.* *Nature Genet.* **25**, 25–29 (2000).
5. Ashburner, M. *et al.* *Genome Res.* (in the press).
6. Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
7. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
8. Kawai, J. *et al.* *Nature* **409**, 685–690 (2001).

♦ <http://www.geneontology.org>

Attempting to standardize gene names across species is pointless.

Misclassification of pest as 'fungus' puts vital research on wrong track

Sir — In your News report on the re-emergence of late blight in Russia¹, the causal agent is referred to as "the fungus *Phytophthora infestans*". The genus name literally means 'plant destroyer' and is derived from the Greek words *phyton* and *phthora*, so the correct spelling is *Phytophthora infestans*.

Phytophthora is grouped in the class of Oomycetes, which behave like fungi in that their main body consists of a network of hyphae that grow at the tip, and they propagate via spores. Yet they are no longer classified in the kingdom Fungi.

Molecular phylogenetic analysis has clearly demonstrated that Oomycetes evolved completely independently from the true fungi — the Ascomycetes and Basidiomycetes — and instead are more closely related to golden-brown algae and heterokont algae in the eukaryotic crown group of Stramenophiles. Lynn Margulis² has classified them in the kingdom Protocista.

Hence, it is debatable whether access to fungicides would help small farmers in Russia to fend off a potato famine. The misconception that *P. infestans* is a fungus has put us on the wrong track for decades. Many fungicide targets are absent in oomycetes, and our knowledge of their biology is limited. New approaches are needed to find novel drug targets and to develop what I suggest should be called 'oomicides'.

Understanding the biology of oomycetes should reveal how these notorious pathogens interact with plants and why host resistance is lost so quickly, providing new leads for durable resistance-breeding strategies.

The aggressive strains currently prevailing in Russia, Western Europe and the United States are the result of a worldwide population displacement of *P. infestans* that started in the mid-1970s. Unlike the earlier infestation discussed in your News report that caused the Irish famine in the 1840s (see this issue, pages 644 and 695), the 1970s infestation included strains with the two different mating types^{3,4}, so *P. infestans* can now propagate sexually in almost every area in the world where potatoes are grown. The egg-shaped sexual spores, oospores, can survive in soil independent of the host plant and can act as an extra inoculum source early in the growing season. Moreover, sexual recombination allows the pathogen to adapt even more easily to adverse conditions.

Fending off a potato famine in Russia requires more than access to suitable pesticides for small farmers. What is really needed is a rational design of control strategies, based on knowledge of the behaviour of the late-blight pathogen and on implementation of this knowledge by pathologists, breeders, biotechnologists and agronomists.

Francine Govers

Laboratory of Phytopathology, Wageningen University, Binnenhaven 9, 6709, PD Wageningen, The Netherlands

1. Nature **410**, 1011 (2001).
2. Margulis, L. & Schwartz, K.V. *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth* (3rd edn) 109–202 (W. H. Freeman, New York, 1998).
3. Drenth, A., Tas, I. C. Q. & Govers, F. Eur. J. Plant Pathol. **100**, 97–107 (1994).
4. Fry, W. E. et al. Ann. Rev. Phytopathol. **30**, 107–130 (1992).

e-access

Nature's debate on access continues at <http://www.nature.com/nature/debates/e-access/index.html>.

Who is prepared to pay, and how much?

Sir — Free or fee, is that the question facing science journal publishers? It shouldn't be. Whatever the method of publishing, it is not without costs. The question should be: who pays the fees and are they reasonable?

As a publishing and information consultant, I find there are two elements that make science journal publishing a somewhat unusual industry. First, journals and the articles published in them are monopolies; no article is published more than once and, once published, an article is available only from the original publisher, although sometimes indirectly (for example via document delivery services). Second, users — authors and readers — do not usually pay for subscriptions or licences; those who do are usually not users. The science journal publishing industry has been likened to the cat food industry: the consumer may be very discerning, but the owner buys the food.

This lack of a 'free market' and its competitive mechanisms means that journal publishing does not follow the usual economic laws of supply and demand. Prices are detached from value or quality, real or perceived, and pricing practices are based on costs plus profit expectations, which are largely at the discretion of the publishers. There is no market correction of prices as a result of competition. Assuming that costs are similar for most publishers, variations in profit expectations cause much of the problem of high prices faced by libraries.

The costs of publishing are growing. For example, the growth in the number of papers published means that even without profit margins, libraries cannot maintain anything like a comprehensive collection

Drug test warning

Sir — Your News article "Researchers strike back in animal-rights row" (*Nature* **411**, 7; 2001) prompts me to ask why drug manufacturers do not put prominent labels on their products comparable to the health warnings on cigarette packets, stating "Development of this drug was made possible by ethically controlled tests on animals"? This would give animal-rights activists and the public a regular opportunity to see the implications of the alternatives available when they come to need, for example, a modern painkiller or chemotherapeutic cancer drug.

T. C. Whitmore

Department of Geography, University of Cambridge, Downing Place, Cambridge CB2 3EN, UK

in their chosen fields, unless unit costs are massively reduced. Those who support the Public Library of Science initiative clearly believe that the costs can be so low that recouping them in only six months would be feasible. They may be right, but profit margins of the kind the industry is used to will be reduced, and publishers will not give up these profits willingly. What is needed is the introduction of true commercial competition with the development of viable new business models. In these, the costs should be kept to a bare minimum, the balance of payments (who pays what proportion of the costs) should be addressed and profit margins should be within a range that most parties would consider fair.

Costs can be kept low by taking full advantage of electronic technology. The balance of payments can be addressed by asking those who arguably benefit most — authors — to pay their share in submission charges. Large profit margins are likely to be reduced in such models in any event.

Peer-reviewed research papers could be accessible via the Internet at very low prices or even free of charge to the reader, and print editions could be made available on demand. BioMed Central (<http://www.biomedcentral.com>) seems to be developing in that direction. It has a good chance of being successful if authors can be persuaded that paying submission charges is in their interest, and if costs can be kept low enough to be covered by those charges. Supplementary revenue from advertising and sponsorship, if sustainable, can lower the thresholds of viability.

More of these models and the science publishing industry might see some real competition!

Johannes Velterop

9 Benfleet Close, Cobham, Surrey KT11 2NR, UK

Diverse origins of biodiversity

A provocative look at how new species arise.

**Frogs, Flies & Dandelions:
The Making of Species**

by Menno Schilthuizen

Oxford University Press: 2001. 245 pp.

£17.95, \$25

Richard G. Harrison

Two fundamental observations about the natural world motivate the work of virtually all evolutionary biologists. The first is the observation of diversity — that natural communities harbour an enormous variety of species. The second is the observation of adaptation, that many or most species appear to exhibit a 'good fit' to their environment. Studies of presumed adaptations are commonplace, as are descriptions of diversity and discussions of how it is maintained. But what of the origin of diversity? Much less has been written about how new species arise — although the process of speciation is central to evolutionary biology.

Frogs, Flies & Dandelions is an engaging and informative introduction to the science of speciation and to the history of the science. As a biologist who has studied speciation for many years, I was fully prepared not to like the book. I imagined that, like much popular science, it would be too simplistic or cute for my taste; and indeed, it occasionally is. But the book was not only surprisingly fun to read but sufficiently provocative to get me thinking hard, and occasionally to get me really angry, mostly at the sophisticated but selective use of data.

Presumably written for a curious and thoughtful but relatively naive audience, the book has enough substance to be of interest to practising evolutionary biologists. The author is a scientist who has also been a science writer. His experience in both disciplines is evident; the book is chatty and informal, but the science is (for the most part) clearly and correctly explained.

Much of the science is revealed through the experiences and observations of individual scientists. These introduce the reader to how science is done, and they show clearly that ideas change over time, that there is a role for speculation and intuition, and



Leaps and bounds: Darwin's sketches of South American frogs and toads helped him mould his theory.

that ultimately the way in which speciation proceeds can only be fully understood by analysing data collected from a large number of different model systems. Many of the vignettes make science and scientific controversies come alive — far more so than if the book had focused only on ideas. Along the way, the author introduces the natural history of birds, flies and fish (and yes, a bit about frogs and dandelions); these are model systems that have contributed to our understanding of speciation.

Speciation events are traditionally classified according to the geographical context in which they occurred. An important distinction is between allopatric and sympatric speciation: the former refers to speciation that occurs when geographically isolated populations diverge and become distinct

species, the latter to speciation or lineage splitting that occurs within a single population.

The essential difference is the extent to which hybridization and gene exchange connect the diverging lineages; allopatric and sympatric speciation are really the ends of a continuum defined by the amount of gene exchange. In the absence of gene exchange (allopatry), populations may diverge simply by chance (when populations are small) or they may diverge in response to (even very small) differences in the way natural selection acts in the isolated populations. When there is substantial gene flow (sympatry), only relatively strong natural selection can drive divergence. This asymmetry accounts for the traditional view, commonly associated with the evolutionary biologist Ernst Mayr, that allopatric divergence is the dominant —



if not exclusive — mode of speciation.

Rebelling against this tradition, *Frogs, Flies & Dandelions* is an unashamed argument for the importance of sympatric speciation. The current ecological context is viewed as playing a decisive role in the diversification of species, whereas a role for past population subdivision, historical bottlenecks in population size and/or founder events is minimized. Given the intended audience, I can appreciate that Menno Schilthuizen felt compelled to provide a decisive resolution of the questions raised in the book, just as a mystery novel needs a good ending. The reader wants to know who committed the crime — or in this case how new species come to be. So we discover at the end that the once-dominant proponents of allopatric speciation have been soundly defeated by the upstart defenders of sympatric speciation.

It is definitely true that we have retreated from the dogmatic certainty of Mayr's views on speciation, in which a few models of allopatric speciation were presented as the only possibilities. Sympatric speciation certainly occurs — and may occur with great regularity in some groups. The *Rhagoletis* story, in which a fly has adapted to a new host tree, moving from hawthorn to apple, is still the best-documented example. But to claim that the "impact of geographical barriers has been trivial rather than paramount" is going too far. Furthermore, we are told that "time and time again, it seems that good old Darwinian natural selection is what speciation is all about".

Adaptationists in the audience will applaud wildly. But although natural selection probably does drive divergence even in most cases of allopatric speciation, intrinsic barriers to gene exchange often arise simply as a fortuitous consequence of two lineages becoming adapted to different environments. Selection does not in these cases play a direct role in speciation.

So I must add a strong cautionary note: don't believe everything you read, and recognize that, had the author chosen a different set of examples, he could have told a very different story. There is a lot missing from this small book; major segments of the speciation literature are completely ignored. There is no discussion of the distinction between pre-zygotic and postzygotic barriers to gene exchange, the former reflecting differences (for example, behavioural differences) that prevent the zygote, or fertilized cell, from forming and the latter a result of reduced viability or fertility of hybrids. We know a substantial amount about the evolution of postzygotic barriers, but such barriers are not discussed in the book. There is also little discussion of recent advances in phylogeography and the evidence it provides for the importance of geographical isolation in diversification and genetic differentiation. Speciation may no longer be, as Darwin

described it, the "mystery of mysteries", but the answers are not as clear-cut as this book would have us believe. ■

Richard G. Harrison is in the Department of Ecology and Evolutionary Biology, Corson Hall, Cornell University, Ithaca, New York 14853, USA.

More on biodiversity

Biodiversity Dynamics: Turnover of Populations, Taxa, and Communities

edited by Michael L. McKinney & James A. Drake
Columbia University Press, \$28, £19 (pbk)

The Diversity of Life, New Edition

by Edward O. Wilson
Penguin, £8.99 (pbk)

Changing times

Perpetual Motion: Transforming Shapes in the Renaissance from da Vinci to Montaigne

by Michel Jeanneret (translated by Nidra Poller)
Johns Hopkins Press: 2001. 408 pp.
\$49.95, £33.50

Joseph W. Dauben

The astonishing idea that motion, transformation and change are impossible goes back to the Greek philosopher Parmenides and his twin assumptions: "what is, is" and "what is not, is not". Atomists such as Democritus asserted that nature was founded on two essential entities — matter and the void. The Pythagoreans insisted that the one unchanging reality of the world was to be identified with number. And Plato took all of this to its most sublime refinement in arguing that the only true knowledge was unchanging, eternal and forever fixed in the ideal realm of platonic forms. In fifteenth-century Italy, such ideas were popularized by a host of Renaissance neoplatonists and eloquently embodied in the serene, seemingly timeless figures depicted by the painter Piero della Francesca.

The sixteenth century in Europe has also been described as a time of "order and perfection expressed in stable, harmonious, closed artistic creations". But in *Perpetual Motion*, Michel Jeanneret refutes such a static, idealized view. He offers a host of examples to prove that, on the contrary, the century was obsessed with the idea of perpetual motion rather than uniform perfection.

Whether or not one accepts Jeanneret's claim that his argument "for a sixteenth century swept up in change and fascinated by genesis and metamorphosis" is in any way new, he offers much stimulating evidence that the century was indeed captivated by the problem of transformation. He cuts

across both arts and sciences, discussing literature, sculpture and architecture, theories of creation and cosmology, biology and geology, the nature of matter and form — showing how all were subject to continuous change.

He considers in detail Leonardo da Vinci's hydrodynamic studies, which included hurricanes, cataclysms and the dynamics of constant turmoil. And the Italian philosopher and astronomer Giordano Bruno figures in a discussion of the infinite Universe and its plurality of worlds. This, Bruno reasoned, must require that motion, too, be correspondingly infinite and unending.

But what of Jeanneret's general thesis? To back up his view of the sixteenth century, he points out that many of the great works of the Renaissance were presented, not as finished products, but as works in progress. Most remarkably of all, the printing-press, which has often been commended for its ability to make exact and permanent reproductions of the scientific eye, proves not to be so stable a medium as we might think. Jeanneret argues, taking examples from the French writer Michel de Montaigne and the literature of the period, that books were meant to be corrected, supplemented and transformed from edition to edition.

Here the sciences can afford excellent examples as well. Copernicus significantly revised his heliocentric theory — that the Sun was the centre of the Universe. He began from



Fluid forms: da Vinci's theory of water flow through the Earth.

SETH JOEL/CORBIS



A century in flux? *Apotheosis of the Renaissance* by the nineteenth-century painter Mihály Munkácsy.

a draft in the *Commentariolus* (1512), which was revised for the first printed notice of his ideas as the *Narratio prima* (1540). This was then significantly augmented with further changes for the official printing of the first edition of *De revolutionibus* in 1543. Similarly, Vesalius's *De humani corporis fabrica libri septem* appeared in the same year, and a second edition with significant revisions appeared in 1555. Surprisingly, Jeanneret does not include these particularly good examples to bolster his overall argument.

He does make his point, dramatically and on a monumental scale, by turning to architecture. A fascinating section of the book is devoted to the city of Rome. Renaissance artists and architects alike explored the city that had been destroyed by Emperor Charles V. A new city was reclaimed, literally, from the ruins of the old as the building blocks scavenged from the Colosseum and the Pantheon were reborn in St Peter's Basilica and the Palazzo Farnese.

Another take on the pervasiveness of change in the Renaissance art of this period is the familiar work of Michelangelo — the image of the sculptor transforming a block of stone into the lifelike image of a polished cardinal or a half-finished slave is a consummate example of transformation. Jeanneret is particularly interested in Michelangelo's many unfinished works. For Jeanneret, this again relates to motion in the sense that a completed work of art is a finality, whereas a work in progress participates in the continuing possibility for further change.

Despite the many detailed examples, one has to ask how seriously Jeanneret expects us to accept his claim that all this attention to change and transformation constitutes a new way of understanding the sixteenth century. By 1500 the stage had been set for a century and more of transformation that would have been inconceivable previously. Think only of the invention of the printing-press, the fall of Constantinople in 1453 and the discovery of the New World by Columbus in 1492. What sealed the fate of Europe and brought on the turmoil of the next few centuries, both spiritually and politically, were the 95 theses drawn up by Martin Luther in Wittenberg in 1517. All this guaranteed that the sixteenth century would be one of conflict and transformation, and of increasingly enlarging horizons.

For the sixteenth century, nothing enlarged those horizons as dramatically as the new heliocentric cosmology. And it is a surprise that Jeanneret does not consider the copernican transformation in thought that turned the Universe itself inside out, pulling the Sun from the heavens and putting it firmly at the centre of the Universe. By removing the Earth from its traditional place and setting it in motion around the Sun, this image alone might have provided Jeanneret with a paradigmatic example of change and motion.

But he admits to shortcomings. He notes that a book devoted to transformation should have included a chapter on alchemy, or perhaps on Edmund Spenser's tale of metamorphosis, *The Faerie Queene*. A chap-

ter on medicine and the Paracelsians — who advocated the medicinal benefits of chemistry in their efforts to heal the sick — would also have been in order, as would mention of Vesalius's assistant Realdus Columbus, who established the pulmonary circulation of the blood in his *De re anatomica libri XV* in 1559. The list of what Jeanneret might have included could easily be extended.

But Jeanneret dismisses such oversights, and, citing the importance in Renaissance literature of taking the reader as collaborator, he asks the same of his own readers — that they think of even more and better examples he might have used to make his point about the centrality of transformation to Renaissance thought. Indeed, the reader may feel like a collaborator, and be stimulated to counter with further examples. As the great humanist Erasmus said, "No book is wrought such that it cannot be made more perfect". And that, after all, is one of the main points Jeanneret wants to make. ■

Joseph W. Dauben is at the Graduate Center, City University of New York, 365 Fifth Avenue, New York, New York 10016-4309, USA.

When baguette meets coffee

Entrer en matière: Les atomes expliquent-ils le monde?
[in French]

by Pablo Jensen

Seuil: 2001. 247 pp. 19.82 euros

Vincent Dusastre

The French term *entrer en matière* can mean either 'getting into matter' or 'introducing a given topic'. Pablo Jensen — a materials physicist at the Université Claude Bernard's Laboratoire de Physique des Matériaux in Lyon — proposes to do both in his book. He aims to take the reader on a journey into the physics of matter and also to give non-scientists an introduction to his research discipline.

The basic idea is to address down-to-earth questions about the physical and chemical principles governing the properties of matter for readers with minimal scientific knowledge. This refreshing approach is directly inspired by an initiative, called the 'Cafés des Sciences', which was started in 1997 by scientists at the Centre National de la Recherche Scientifique in Lyon, and now extends to the rest of France. Its aim is to improve the dialogue between science and society.

Entrer en matière is particularly accessible because its various sections can be read independently, depending on your mood and your knowledge of the subject. Jensen lucidly portrays the developments and intellectual challenges encountered by physicists in their

attempts to describe and explain the complex and diverse behaviour of matter, and the challenges that remain. And he illustrates his story with practical examples that have economic, environmental, historical and even artistic applications.

The book starts by posing basic questions about the different states of matter — solid, liquid and gas — encountered in everyday life. Or, at least, in the life of a typical French person, for most of these practical interrogations revolve around what goes on in the kitchen or at the table — from what happens when a buttered baguette is dipped into coffee to the structure of meringue. It is not surprising that the cultural obsession with matters culinary has recently resulted in the creation of a research laboratory, at the Collège de France in Paris, dedicated to 'molecular gastronomy'.

Jensen next explains how physicists have tried, with varying degrees of success, to explain the intriguing behaviour of everyday items by 'reconstructing' them from their basic components (the atoms). Toothpaste, water, glass and chocolate all undergo complex and schizophrenic transformations — oscillating between one state and another — depending on external conditions. And Jensen shows how the explanations physicists came up with have had a profound philosophical influence on our current view of matter.

Even today, one aspect of the philosophical debate is whether this 'reductionist' idea of matter is needed to understand the everyday world. Jensen's view is that the fundamental laws governing the interaction between particles at the most intimate level (inside atoms) are not necessarily relevant when explaining the macroscopic world. And he rightly argues that, for any specific example, each level of complexity should be treated separately — for instance, we should not systematically try to explain the macroscopic properties of a material in terms of its quantum mechanics. Although certainly logical in the physical sense, such controversial ideas are not particularly constructive — and are probably not in line with the open spirit of the Cafés des Sciences. They will surely raise a few eyebrows, especially among the particle-physics community.

Jensen then goes on to describe the tricks and tools that physicists use to infer the macroscopic behaviour of a physical system from the supposedly well-known microscopic behaviour of its components. And he gives everyday examples of the physical properties (mechanical, optical, electrical) shared by all materials. He also shows how certain classes of familiar materials — such as steel, plastics, foams and adhesives — have influenced the development of materials science and puts current research interests in context.

The final part of the book, which I found by far the most entertaining and instructive,

offers a superb guide to what the author and his fellow physicists actually get up to in their mysterious laboratories. It reveals physicists as normal human beings, and explains why their search to understand matter makes them hide away from the outside world with their sophisticated and expensive pieces of equipment. To the general public, physicists can seem like alien beings, wasting extravagant amounts of taxpayers' money on useless experiments. This part of *Entrer en matière* should make such readers think again.

For a book that aspires to familiarize the person in the street with the latest developments in science, it's a pity the presentation is so dull — the only colour pictures are on the cover. This may make it less attractive to the wider audience that its accessible writing deserves.

I warmly recommend *Entrer en matière* to anyone (but especially physicists) even remotely interested in discovering more about matter. The book would serve as a perfect introduction to a first-year physics course — especially in French universities, where the material covered is more often about theory than the understanding of physical concepts. I only wish that such a book had been available in my student days.

I hope that this enlightening book, written by someone with such enthusiasm for communicating his knowledge, will be translated into other languages. It could then fulfil its ultimate aim of promoting interaction, communication and debate between scientists and the rest of the world. ■

Vincent Dusastre is an associate editor at Nature.



Twilight creatures

These polar dinosaurs (reconstructed), dubbed *Leaellynasaura amicagraphica*, lived between 100 and 120 million years ago in Australia, which at that time lay farther south than today, joined to Antarctica. *Dinosaurs of Darkness* by Thomas H. Rich and Patricia

Vickers-Rich (Indiana University Press, \$35, £24.95) describes the search for and analysis of fossils that would elucidate how these animals were adapted to cope with the harsh conditions and the poor light that must have existed for much of the year.

Sulphur and holy water

Did a devout clergyman inspire the fictional archetype of the diabolical scientist?

Paolo Mazzarello

"Professor Spallanzani walked slowly through the empty room, his footsteps giving a hollow echo; and his figure had, as the flickering shadows played about him, a ghastly, awful appearance."

This is the diabolical image of the scientist "of Italian origin" in Ernst Theodor Amadeus Hoffmann's tale *Der Sandmann*, written in November 1815. Spallanzani creates a bewitching robot called Olimpia, that attracts the young man Nathanael. Under Spallanzani's enigmatic gaze, Nathanael is captured by an ambiguous witchcraft and falls in love with the mechanical puppet. He goes insane when Olimpia suddenly reveals 'her' real nature.

The wily inventor Spallanzani and his creature Olimpia later appear in the Jacques Offenbach opera *The Tales of Hoffmann*, (libretto by Jules Barbier) performed for the first time on 10 February 1881 at the Opéra-Comique in Paris. In this opera, Spallanzani is close to financial ruin and hopes to make a fortune from his automaton, and the protagonist who falls in love with the robot is now a student named Hoffmann.

I believe the fictional scientist Spallanzani was based on his near-namesake Lazzaro Spallanzani (1729–1799), a Roman Catholic priest who was professor of natural history at the University of Pavia and a member of the Royal Society of London. But where Spallanzani gives his mechanical doll a lifelike appearance in *Der Sandmann*, the Pavian scientist obtained what he called a "real resurrection after death" of small desiccated animals (Rotifera and Tardigrada) by means of subsequent humidification.

These experiments posed disquieting philosophical and theological problems as, according to Aristotle, all living beings were endowed with a soul. In a letter to the French

philosopher Voltaire (quoted in Spallanzani's *Carteggi*), Spallanzani wrote: "What about the problem of the soul? When they are dead, what happens to their souls? During the time in which these animals are dead where do they go? Could we believe that every time they are dead the souls leave their bodies and subsequently go back during resurrection? ... Even if we admit the existence of a soul in these portentous animals, the naturalist and the philosopher are both embarrassed. And if we try to get out of the problem by admitting that these animals do not have a soul, why can't we say the same of many others?"

Strange questions for a naturalist who is also a priest. But stranger still if one considers that the addressee was a well-known scourge of ecclesiastical hierarchies. Voltaire replied: "You are considered the best observer in Europe. ... When a man like you announces that he has brought the dead back to life we have to believe him. ... If there is someone, Sir, that has the right to explain this mystery, this person is you." That mystery, however, remained unanswered at the time.

But Spallanzani did find the answers to other natural mysteries. He disproved the doctrine that microscopic life could arise from spontaneous generation, a theory curiously advocated by another Roman Catholic clergyman and member of the Royal Society, John Turberville Needham. The two priests engaged in an intellectual battle that lasted many years and ended with Spallanzani's views prevailing.

A fanatical researcher, Spallanzani was always trying to understand the most intimate phenomena of nature, and he carried out the most audacious experiments. Keeping a female poodle segregated until it was on heat, he artificially inseminated it with semen "spontaneously" obtained from a male. Cutting the head off a snail, he observed its capacity to regenerate. Blinding bats, he discovered that their ability to orientate themselves did not diminish and thus suspected the existence of a new sense in these animals.

To dissociate the mechanical from the chemical phenomenon of digestion, he swallowed little perforated wooden tubes or small cloth sacks containing bread or meat. Moreover, using gastric juices obtained from dissected animals and from himself by induced vomiting, he achieved an artificial digestion of meat and vegetables, thus disproving the prevailing idea of the



Faustian act: inside Spallanzani's laboratory during a performance of *The Tales of Hoffmann*.

existence of a "vital force" in the process.

Spallanzani was the first to observe blood leukocytes, to prove the animal nature of marine sponge, to demonstrate the respiratory process in many tissues (not only in the lungs, as proposed by Lavoisier). He was a polymath who made original observations in many other fields of science, from mineralogy to vulcanology, from chemistry to oceanography.

It was no surprise that this priest was perceived, for many years after his death, as a mixture of sulphur and holy water, a legendary wizard of experimentation. His successful programme of research exploring the extreme philosophical boundaries of eighteenth-century science made him a literary symbol of scientific astuteness in Hoffmann's tale.

The history of literature contains various archetypal figures of scientists. Leonardo da Vinci is the icon of the universal genius, Galileo the scientist in conflict with power (see, for example, Bertolt Brecht's *Life of Galileo*), Einstein the genial eccentric scientist, Archimedes the personification of the power of mathematical thought. Spallanzani, with his fanatical thirst for knowledge, certainly enters this gallery as the wily Faustian scientist able to investigate and dominate the most elusive laws of nature. ■

Paolo Mazzarello is at Volta College, University of Pavia and IGBE – CNR, Via Abbiategrasso 207, I-27100 Pavia, Italy.

FURTHER READING

Spallanzani, L. *Opuscoli di Fisica Animale e Vegetabile* (Società Tipografica, Modena, 1776).

Spallanzani, L. *Carteggi* (ed. Di Pietro, P.) Vol. 11 (Mucchi, Modena, 1989).

Spallanzani was a polymath who made original observations in many fields of science, from chemistry to oceanography.

Physics in the noise

Michael F. Shlesinger

Random walks span a gamut from watching winnings fluctuate in games of chance, to simple brownian diffusion, to modern studies of nonlinear dynamics. As with many important concepts in twentieth-century physics, random walks had a boost from Einstein who, in 1905, was focusing on brownian motion. In 1785, Jan Ingenhauz had placed charcoal powder on an alcohol film and saw the grains move randomly; in 1828, the botanist Robert Brown had reported erratic dancing of small particles in fluids at rest. Einstein considered these fluids to be composed of discrete molecules whose many collisions with a brownian particle caused them to jump in random directions—a random walk. His analysis not only explained brownian motion, but also bolstered the case for the existence of atoms (which at that time was not universally accepted).

Einstein's approach was founded on the principal that noise, like other natural phenomena, is an expression of physical law. He showed that the brownian particle obeys the diffusion equation; his formula for the mean-squared displacement involved Avogadro's number as a factor in the diffusion constant. Jean Perrin's careful observations of brownian motion allowed him to calculate Avogadro's number in this manner and hence to win the 1926 Nobel prize.

As is often the case, precursors were discovered once this work was recognized as important. Lord Rayleigh, in 1880, made the connection between diffusive heat flow and random scattering. In 1900, Louis Bachelier introduced the random-walk theory of market fluctuations, finding that bond prices could diffuse in the same manner as heat.

And back in the 1700s, Jacob Bernoulli and Abraham DeMoivre studied the random walk of a gambler's stake as it changed with the outcome of a fair game of chance.

All these cases, including Einstein's, describe a sequence of uncorrelated random changes of a quantity (position, price or winnings). Even though each change is random, a probability distribution can be derived, allowing the formulation of equations that predict the probability of future behaviour. The underlying equation is the diffusion equation; its solution is the bell-shaped gaussian curve, which widens linearly with time (in these examples, with the number of collisions, market trades or games played).

But random walks can also display complex behaviour that cannot be described by brownian motion. A famous example is the experiments on glassy semiconductors used in photocopying in 1974 at the Xerox labs. A flash of light creates charged pairs of electrons and holes, the movements of which are monitored by the current they generate. Harvey Scher and Elliott Montroll realized that the charges were moving between traps, and that the distribution of trapping times was so wide that the average time to get out of a trap exceeded the duration of the experiment. In contrast to the brownian model, the number of jumps of a random-walking charge only grew at the 0.45 power of time, instead of linearly with time. The resulting probability distribution was more peaked than a typical gaussian, with a width that grew more slowly than linearly with time. These results broke the standard dimensional analysis of diffusion theory and led to unusual scaling laws. The time, T , for a charge to cross a sample of length L in an electric field, E , scales as (L/E) to a strange power, 2.2, rather than one, as would be found for a classical random walk. More

recently, this type of random walk has been applied to defect motion in glassy materials to explain slow 'stretched exponential' relaxation near the glass transition temperature.

Turbulent diffusion is another example of a non-brownian random walk. In 1926, Lewis Fry Richardson found that the second moment of the probability distribution for the relative separation of two

Random walks

This rich and varied phenomenon is involved in particle trajectories that exhibit fractal, chaotic and turbulent behaviour.

particles undergoing turbulent diffusion grows as time cubed. Paul Levy, in the 1920s, obtained probability distributions for a class of random walks with infinite second moments of the jump-length probability. These non-gaussian 'Levy flights' do not possess the smooth flow of a diffusion process. Because of the likelihood of longer and longer jumps, the flight paths burst out from their origin, hitting a fractal clustered set of points. But how could such mathematical conjuring ever find a physical application? Levy's work stayed in the mathematics literature, unknown in physics until recently.

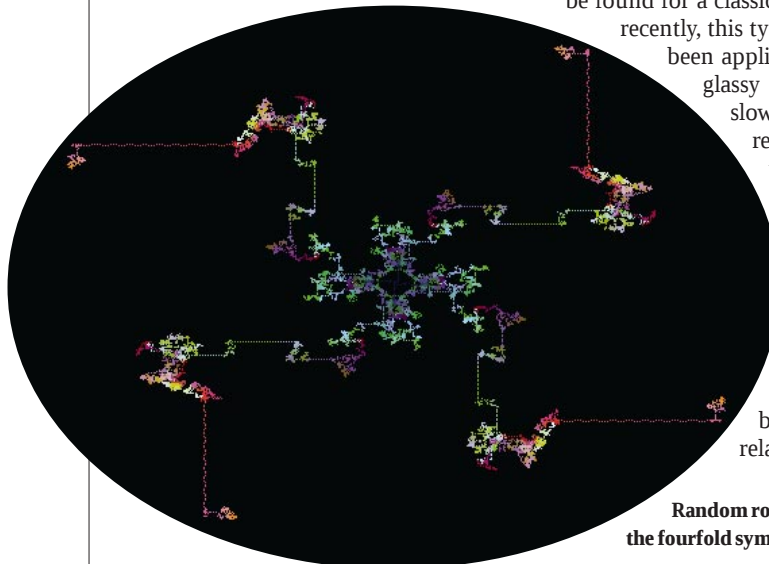
A physicist does not ask how large is a random-walk jump, but rather how far the jump has proceeded in a time t . This introduces the velocity, V , into the random walk. Even if the random walk is in the middle of an infinite jump, à la Levy, the walker will only have travelled Vt within a time t . Thus, the infinity never enters the calculation. The process is called a Levy walk or drive, and its non-gaussian probabilities describe supra-diffusion with second moments growing as a power of time greater than one. Turbulent diffusion falls into this category.

It was a big surprise that random Levy walks appeared in dynamic systems that are deterministic without a hint of probability in the equations of motion. The key is that seemingly random-like behaviour can arise through nonlinearity. This is the famous chaos paradigm. In certain nonlinear systems, trajectories can diffuse in a chaotic sea, alternating with becoming trapped for a wide range of time scales on almost closed, nested, nonlinear orbits. These systems have nonlinear sensitivities that produce trajectories reminiscent of Levy walks. Examples include the standard map of a periodically kicked rotor, the Zaslavsky web map of a periodically kicked oscillator, and Geisel's map for chaotic-phase rotation in a Josephson junction. ■

Michael F. Shlesinger is in the Physical Sciences Division, Office of Naval Research, 800 North Quincy Street, Arlington, Virginia 22217, USA.

FURTHER READING

Hughes, B. D. *Random Walks and Random Environments* (Clarendon, Oxford, 1995).
Weiss, G. H. *Aspects and Applications of the Random Walk* (North-Holland, Amsterdam, 1994).
Shlesinger, M. F. in *Encyclopedia of Applied Physics* Vol. 16, 45–70 (VCH, Berlin, 1996).
Klafter, J. et al. *Phys. Today* **49**, 33–39 (1996).



Random route: a Levy walk made by the fourfold symmetric Zaslavsky map.

Fifty years of inactivation

Richard W. Aldrich

Potassium channels can be closed by a process known as inactivation — this is, for instance, how nerve cells regulate firing frequency. Events involved in inactivation are now revealed in unprecedented detail.

The ion channels that span cell membranes and moderate the transmembrane flow of electrical currents regulate a whole range of biological phenomena. The channels themselves are controlled in a variety of ways, one notable group being 'voltage-gated' — they open in response to, and then cause, changes in membrane voltage. Almost 50 years ago, the Nobel laureates Alan Hodgkin and Andrew Huxley showed the physiological importance of the additional process of channel inactivation, but pointed out that "Details of the mechanism will probably not be settled for some time"¹. Judging from the paper by MacKinnon and colleagues on page 657 of this issue², it seems that time has now come.

Inactivation has been extensively studied in two of the main types of ion channel, those that allow the passage of sodium or potassium ions. In sodium channels, inactivation is important for the termination of the nerve impulse and for setting the period before the nerve cell can fire again; in potassium channels, it shapes the electrical signalling properties of many types of nerve, muscle and other cells. Since Hodgkin and Huxley's work, the application of increasingly sophisticated techniques has led to a progressively greater understanding of inactivation (which appears to occur in similar ways in both types of channel). Decades of hard work, by electrophysiologists in particular, have now culminated in the paper from MacKinnon's group.

So, how did we get here? It was the size and simplicity of the famous squid giant axon that allowed Hodgkin and Huxley to describe the time- and voltage-dependence of the channels that underlie the nerve impulse. Technical advances with 'voltage-clamp' configurations led to the investigation of cells with more complex electrical properties, and the discovery of a wide variety of new channel behaviours — including potassium-channel inactivation³.

It was subsequently found that inactivation could be removed by intracellular enzymes⁴, could be mimicked with channel-blocking agents such as derivatives of tetraethylammonium⁵, and had little intrinsic voltage dependence^{6,7}. These results were unified with the development of the 'ball-and-chain' model in the late 1970s⁸. This model has a positively charged inactivation particle (the ball), on a tether (the chain),

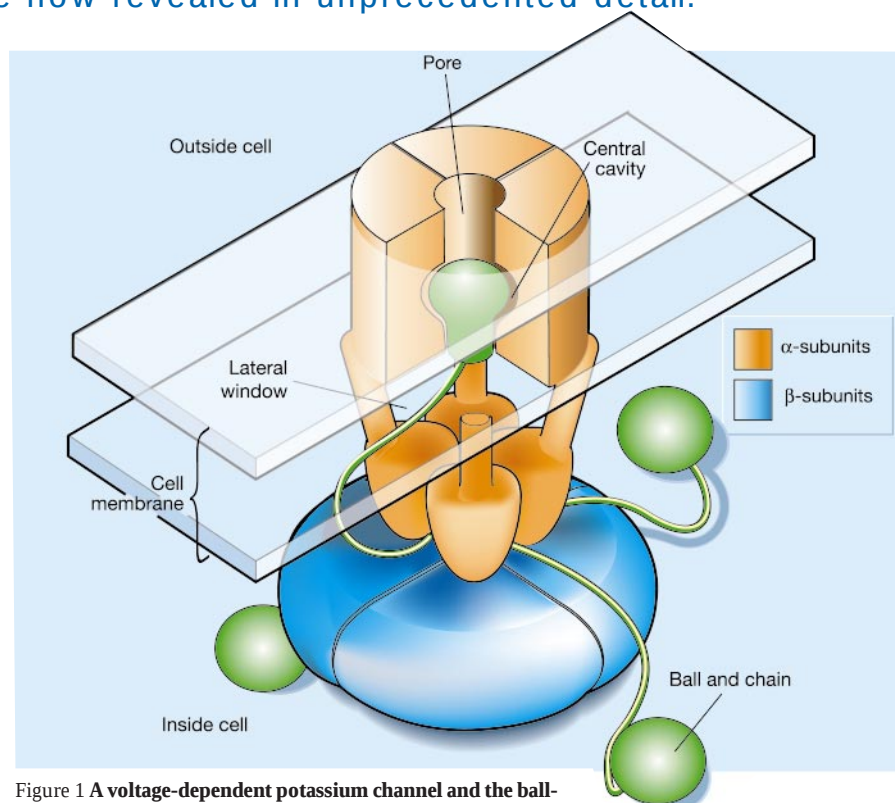


Figure 1 A voltage-dependent potassium channel and the ball-and-chain model of channel inactivation. MacKinnon and colleagues² (who refer to the system as the N-terminal inactivation gate) provide a wonderfully detailed view of events during inactivation. The channel is composed of four α -subunits, which span the cell membrane, and four β -subunits that lie just inside the membrane. The α -subunits have three components, the uppermost being the pore-forming domain (one part of which is not shown here so as to reveal the pore). The ball and chain of the eponymous theory of channel inactivation are shown in green. For inactivation to occur, a positively charged inactivation particle (ball) has to pass through one of the lateral windows and bind in the hydrophobic binding pocket of the pore's central cavity. This blocks the flow of potassium ions through the pore. There are four balls and chains to each channel, but only one is needed for inactivation.

which can block the channel by binding to it (Fig. 1).

The advent of molecular cloning added flesh to the bones of this model. An archetypal potassium channel is the Shaker channel in the fruitfly *Drosophila*. The discovery of different inactivation rates among Shaker variants⁸ led to the use of site-specific mutations to locate the inactivation particle to the first 20 amino acids residues at the N (amino)-terminal end of the channel protein, and the chain to the 50 or so residues next to it⁹.

Further advances then followed from experiments with a synthetic peptide having the same sequence as the inactivation particle. This approach allowed characterization of the particle in molecular terms, and detailed mechanistic studies of its inter-

action with the binding site in the channel pore¹⁰. These studies revealed that the molecular determinants of an effective inactivation particle were rather vague; the first ten or so amino acids had to be hydrophobic, and the next ten or so needed to have a net positive charge. Altering the hydrophobic part affected the unbinding rate, and altering the charged part affected the binding rate¹¹. These findings confirmed proposals that the ball, like hydrophobic tetraethylammonium derivatives, interacted with a hydrophobic region of the channel, and that electrostatic attraction played a role in getting the ball to its binding site^{5,6}.

Several results pointed to a complex interaction between the inactivation particle and its binding site. The binding rate was

very slow and the temperature dependence very high; and amino-acid position in the sequence was not particularly important as long as the sequence conformed to the general requirements of a hydrophobic region followed by a region of net positive charge^{11,12}. These findings led to the proposal that the inactivation domain was quite flexible but could bind in only some of its conformations¹¹.

So far so good. But a surprise then came from structural studies — they showed that the route to the inner mouth of the channel was not along the central axis but through four lateral 'windows' between the cytoplasmic domains of the channel's four α -subunits^{13–15} (Fig. 1). This meant that the blocking ball would have to enter through these windows and then work its way up to the narrow part of the channel mouth. Such a complex conformational change seemed so unlikely that, for some, the whole ball-and-chain idea seemed doubtful.

The new results from MacKinnon and colleagues² settle the issue. Inactivation does work through the ball-and-chain principle. And our understanding of it has been brought to a structural level that unifies the earlier biophysical conclusions (see the authors' Fig. 4 on page 659). The paper² presents three main findings.

First, the authors identified the amino-acid residues in the channel pore that interact with those of the inactivation particle. The technique they used involves systematic mutation of both pore and inactivation peptides, and measurement of changes in binding energies to determine specific interactions between residues. MacKinnon and co-workers found interactions between hydrophobic residues of the inactivation peptide and residues deep within the central cavity of the channel pore, showing that the inactivation ball has access to a binding site in the central cavity, even though it must get there by way of the side windows.

Second, the authors have mapped the position of these interacting residues onto the known structure of a bacterial potassium channel. The result indicates that, in the inactivated state, the inactivation particle becomes elongated as it snakes up into the pore. There are no secondary structural features, such as α -helices, in the bound conformation. This fits with the earlier, less direct, electrophysiological studies suggesting that significant shape changes are involved in binding, and pointing to the existence of a hydrophobic binding pocket^{5,11,16}.

Finally, there are the results from structural studies using the inactivation mimic, tetrabutylantimony, an analogue of tetrabutylammonium. The tetrabutylantimony can be seen in the crystal structure, binding in the central cavity, right next to the residues that interact with the inactivating peptide. This confirms in exquisite detail Clay Arm-

strong's proposal of about 30 years ago⁵ that inactivation works by a mechanism similar to that in which tetraethylammonium derivatives block channel function.

These new findings² provide a detailed molecular view of channel inactivation that fits satisfyingly well with the results of a variety of techniques over the past half-century. This is a point worth stressing. Understanding the mechanisms of biological processes depends on technical advances, and on multidisciplinary approaches. Both principles are beautifully exemplified by the work of MacKinnon and colleagues. ■

Richard W. Aldrich is in the Department of Molecular and Cellular Physiology, and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA.

raldrich@leland.stanford.edu

1. Hodgkin, A. L. & Huxley, A. F. *J. Physiol. (Lond.)* **117**, 500–544 (1952).

2. Zhou, M., Morais-Cabral, J. H., Mann, S. & MacKinnon, R. *Nature* **411**, 657–661 (2001).
3. Hille, B. *Ionic Channels in Excitable Cells* 2nd edn (Sinauer, Sunderland, MA, 1992).
4. Armstrong, C. M., Bezanilla, F. & Rojas, E. *J. Gen. Physiol.* **62**, 375–391 (1973).
5. Armstrong, C. M. *J. Gen. Physiol.* **54**, 553–575 (1969).
6. Armstrong, C. M. & Bezanilla, F. *J. Gen. Physiol.* **70**, 567–590 (1977).
7. Aldrich, R. W., Corey, D. P. & Stevens, C. F. *Nature* **306**, 436–441 (1983).
8. Timpe, L. C., Jan, Y. N. & Jan, L. Y. *Neuron* **1**, 659–667 (1988).
9. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. *Science* **250**, 533–538 (1990).
10. Zagotta, W. N., Hoshi, T. & Aldrich, R. W. *Science* **250**, 568–571 (1990).
11. Murrell-Lagnado, R. D. & Aldrich, R. W. *J. Gen. Physiol.* **102**, 977–1003 (1993).
12. Nobile, M., Olcese, R., Toro, L. & Stefani, E. *Exp. Brain Res.* **114**, 138–142 (1997).
13. Kreusch, A., Pfaffinger, P. J., Stevens, C. F. & Choe, S. *Nature* **392**, 945–948 (1998).
14. Gulbis, J. M., Zhou, M., Mann, S. & MacKinnon, R. *Science* **289**, 123–127 (2000).
15. Sokolova, O., Kolmakova-Partensky, L. & Grigorieff, N. *Structure* **9**, 215–220 (2001).
16. Holmgren, M., Smith, P. L. & Yellen, G. *J. Gen. Physiol.* **109**, 527–535 (1997).

Astronomy

A twisted look at the X-ray sky

Webster Cash

By adopting technology developed for particle-physics experiments, astronomers have created a new X-ray tool for probing exotic parts of the Universe.

In astronomy there are basically only four types of observation that can be made using electromagnetic radiation (photons). These involve measuring the photon

direction (imaging), their energy and frequency (spectroscopy), their polarization (polarimetry), and counting the numbers of photons (photometry). These techniques

Plant pathology

Reverend Berkeley's devil

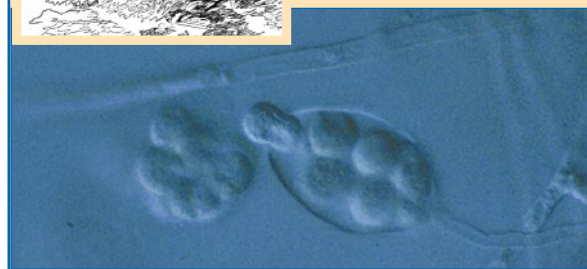


Organisms known as oomycetes or water moulds sound innocuous enough. There are 700 or so species of them, and although they are referred to as fungi (as the 'mycetes' in their name implies), they are more closely related to kelps and diatoms than mushrooms. But these microorganisms can have global effects. Earlier this year, for instance, one species — *Saprolegnia ferax* — was implicated in the widespread

decline in amphibian populations (J. M. Kiesecker *et al.*, *Nature* **410**, 681–684; 2001).

The most infamous member of the group is *Phytophthora infestans*, the agent of late blight of potato. In 1845 and 1846 it devastated the potato crop in Ireland, and one million people starved. The historical image here portrays the struggle for existence in the aftermath of the famine. The micrograph shows the infectious *Phytophthora* spores being released from structures known as sporangia.

By identifying a fungus as the cause of the epidemic, the Reverend Miles Berkeley broke ranks with his fellow clergy who blamed the devil. But where



provide complementary information and so are vital for exploring all wavelengths of radiation. So far astronomers have not been able to efficiently detect polarization of photons at X-ray wavelengths, but a new instrument described on page 662 of this issue¹ by Enrico Costa, Ronaldo Bellazzini and their co-workers in Rome and Pisa could change all that.

X-ray astronomy has revealed some of the most violent and compact spots in the Universe, such as the surfaces of pulsars, the close orbits around giant black holes, and the blast waves of supernova explosions. By making efficient use of the few photons emitted by disks around black holes and other objects, X-ray astronomers have successfully applied photometry, imaging and spectroscopy to these hot, energetic and often variable sources. But polarimetry is largely ignored at X-ray wavelengths because of the inefficiency of existing instruments. The new X-ray polarimeter developed by the teams in Rome and Pisa promises to revolutionize space-based observations.

The first and only generally accepted measurement of polarized X-rays from an astronomical object was taken more than a quarter of a century ago when a Bragg crystal polarimeter in orbit around the Earth was used to observe the Crab nebula². This nebula is a remnant of the blast wave from a supernova that expanded into the surrounding interstellar gas, and is unusual because it has a bright pulsar at its core. Spinning at about 30 times per second, the intense magnetic field of this neutron star creates a dynamic nursery for high-energy particles and X-rays. Figure 1 shows an X-ray image

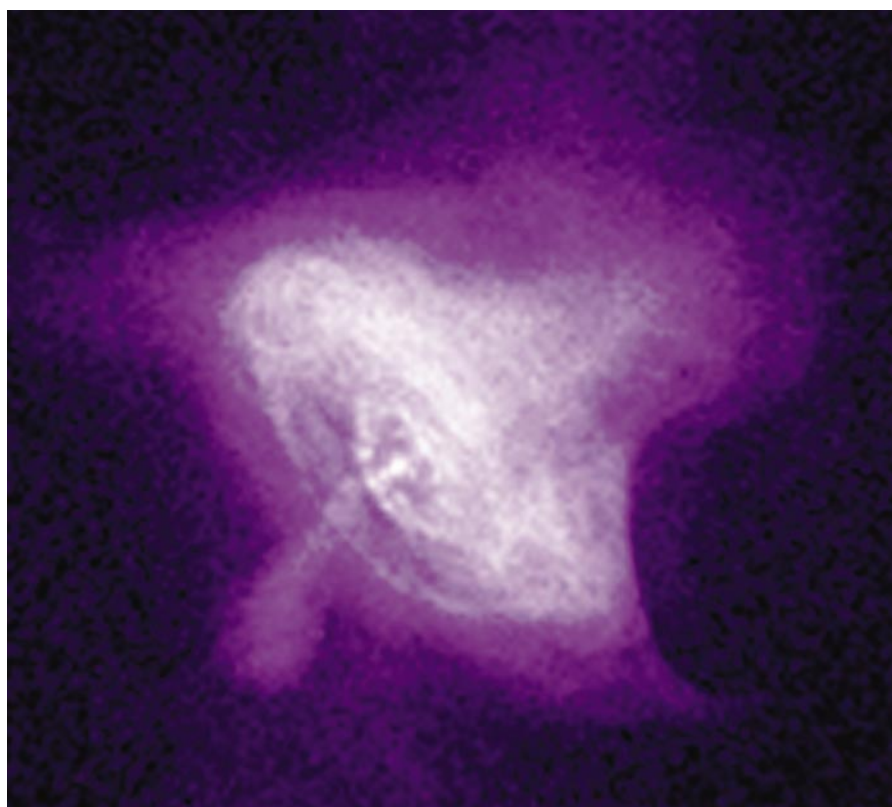


Figure 1 Image of the Crab nebula captured by the Chandra X-ray observatory. The X-ray emission forms a torus and jets, which emanate from the central pulsar. The Crab is the only X-ray source with proven polarization, which probably comes from synchrotron emission in the torus. A new efficient X-ray polarimeter developed by Costa *et al.*¹ could tell us more about the geometry of this and other sources of X-ray radiation.

of the centre of the Crab nebula taken by the Chandra X-ray Observatory³ in space. Right in the centre of the nebula is a bubble

of highly intense, polarized radiation known as synchrotron radiation that was formed by high-energy particles flung from the pulsar and trapped in the surrounding magnetic field. Wrapped around the central pulsar is a torus of synchrotron radiation, and jets of radiation appear to emerge from the poles of the pulsar.

When high-energy electrons are caught in a magnetic field they emit synchrotron radiation caused by acceleration due to the curvature of their paths. As with other electromagnetic waves, the vibrating electric and magnetic fields that make up X-rays can point in different directions, known as polarization states. The degree of polarization can vary from 0%, when the orientation of the electric fields in the photons is random, to 100%, when all the electric fields point in the same direction, known as the polarization angle. In general, synchrotron radiation is elliptically polarized, but a component of the radiation is linearly polarized⁴. Synchrotron emission is the source of the 19% polarization detected in the Crab nebula. Being able to measure the polarization of the radiation as a function of its position across the Crab nebula would lead to a much better picture of the geometry of the magnetic field in the nebula and its central pulsar.

There has been no unambiguous detection of X-ray polarization from a celestial

had the disease come from? In this issue, Jean Ristaino and colleagues describe how they have used modern methods to tackle the question (*Nature* **411**, 695–697; 2001).

In the late 1980s and early 1990s, fungicide-resistant strains of *P. infestans* emerged, and potato blight is now recognized as a significant present-day menace to agriculture. Much of the genetic diversity in the species is found in the Toluca Valley of Mexico, where a variety of strains reproduce sexually. Elsewhere, until recently, most plant infections were caused by a single strain called US-1, which reproduces asexually and has spawned a single, extended clonal lineage. Given its contemporary distribution, plant pathologists believed that

US-1 was a direct descendant of the fungus responsible for the Irish potato famine.

In a remarkable piece of molecular detective work, Ristaino and colleagues have isolated and analysed *Phytophthora* DNA from herbarium specimens of infected leaves collected during and after the famine. Mutations in the mitochondrial DNA of the pathogen serve as markers that distinguish between four of the most common clones. Amplification of these sequences from fragments of the dried leaves confirms Berkeley's identification, but shows that the Irish famine was not caused by US-1. This study also casts doubt on the idea that the fungus spread from Mexico, and shows that further genetic analysis will be

needed to work out its recent evolutionary history. Locating the centre of origin of *P. infestans* should help in identifying resistant varieties of host plants and genes that might confer protection against it. As the ancestral home of the potato, South America may be a fruitful hunting ground.

The new findings mean that some ideas about the origin of historical plant disease epidemics will need to be re-evaluated. But this application of molecular archaeology should hearten herbarium directors who are fighting to preserve their treasured archives.

Nicholas P. Money is in the Department of Botany, Miami University, Oxford, Ohio 45056, USA.
e-mail: moneynp@muohio.edu

source since the Crab discovery — all other sources are either too faint or are largely unpolarized. To capture the polarization of these faint sources would need a device capable of measuring the polarization angle of every photon collected by an X-ray telescope. Through the joint efforts of the two Italian teams¹ we now have such an instrument. It functions mainly as a photon-counting detector, its overall architecture being similar to a radiation Geiger counter or a proportional counter used in particle physics.

In the new device, an X-ray entering the chamber interacts with the gas inside, emitting an energetic photoelectron. The polarimeter then measures not only the presence of the electron, but also the path it has taken. The motion of the electron is driven by the direction of the electromagnetic field in the original photon, thereby recording the linear polarization of the X-ray. When enough events have been detected, the polarization of the emitting source will have been measured. This device makes maximal use of the available information, and, when placed at the focus of a large X-ray telescope in orbit, will be able to detect as little as 1% polarization in sources a thousand times fainter than the Crab Nebula. This will open a new window on the geometry of X-ray-emitting sources.

Synchrotron radiation is not the only source of polarized radiation in space. Equally important is the scattering of X-rays. When an X-ray encounters a free electron, it can change direction, imparting some of its energy to the electron through a process known as Compton scattering. The radiation that emerges is linearly polarized, with its electric field perpendicular to the plane that contains both the incident and the scattered photons. The degree of polarization is a function of the angle through which the X-ray is scattered, reaching 100% for a deflection of 90°. This effect can be used to determine the basic geometry of a number of X-ray sources.

For a start, the X-ray-emitting regions of active galactic nuclei and quasars demand a closer look. The hot gas swirling into the giant black hole that powers these sources emits copious X-rays through both thermal and non-thermal processes. Narrow jets of particles travelling very close to the speed of light can form by means that are only partially understood. The X-ray spectrum that emerges is complex, featuring emission from several parts of the source⁵. Radiation passing close to the black hole will be curved, and its polarization twisted. Furthermore, radiation from one part of the source can scatter off another, creating more features in the spectrum⁶. Measurement of polarization as a function of the energy of the X-rays can tell us a lot about whether the radiation has been scattered, and through what angles. This

will be a unique test of the physics of these fascinating objects.

Finally, significant improvements in observing power usually lead to important new discoveries. Perhaps when Costa, Bellazzini and their colleagues loft the first of their efficient polarimeters above the atmosphere they will be able to detect polarization in unexpected places — maybe from the surfaces of thermal neutron stars, or even from interstellar shocks that arise when high-speed plasma collides with quieter regions of cooler gas. We can all hope that this new device will follow the established pattern of

invention and discovery, making observations that will change our understanding of the cosmos.

Webster Cash is in the Department of Astrophysical and Planetary Sciences of the University of Colorado, Boulder, Colorado 80309, USA.
e-mail: cash@casa.colorado.edu

1. Costa, E. *et al.* *Nature* **411**, 662–665 (2001).
2. Weisskopf, M. C. *et al.* *Astrophys. J.* **208**, L125–L128 (1976).
3. <http://chandra.harvard.edu>
4. Rybicki, G. B. & Lightman, A. P. *Radiative Processes in Astrophysics* (Wiley, Chichester, 1979).
5. Mushotzky, R. F., Done, C. & Pounds, K. A. *Annu. Rev. Astron. Astrophys.* **31**, 717–761 (1993).
6. Reynolds, C. S., Young, A. J., Begelman, M. C. & Fabian, A. C. *Astrophys. J.* **514**, 164–179 (1999).

Cancer

Telomerase meets its mismatch

Raju Kucherlapati and Ronald A. DePinho

Cells cannot survive without telomeres, the sequences that cap the ends of chromosomes, so cancer cells activate a telomere-generating enzyme. Studies of yeast now hint that they have a second way to make telomeres.

A human cell contains nearly six billion base pairs of DNA. With these huge numbers, it is almost inevitable that, during the lifetime of the cell, the DNA will accumulate errors, from either environmental influences or mistakes in DNA replication. To fix this damage to their genomes, all organisms have caretaker proteins, including those of the DNA 'mismatch-repair' system¹. As well as mending damaged DNA, mismatch-repair proteins also prevent slightly dissimilar DNA strands from swapping chunks of sequence (recombining). So it is not surprising that genomes become rather unstable when their mismatch-repair genes are mutated and no longer function. As Rizki and Lundblad² show on page 713 of this issue, mutations in mismatch-repair genes also allow yeast cells to survive in the absence of an enzyme called telomerase — a result that has implications for human cancer.

Because mismatch-repair proteins have such a crucial role in mending damaged DNA, errors accumulate more rapidly when the genes encoding these proteins are mutated, predisposing humans to several different cancers. In particular, defects in one subset of mismatch-repair genes result in an inherited cancer syndrome called hereditary non-polyposis colorectal cancer (HNPCC)³. A good model to explain how the loss of mismatch-repair genes leads to an increased susceptibility to cancer is that, as errors accumulate, some will inactivate tumour-suppressing genes or activate cancer-promoting genes (oncogenes). In this way, cells would reach a critical threshold of tumour-promoting errors. This is one aspect of the genomic instability that results from defects in mismatch-repair proteins.

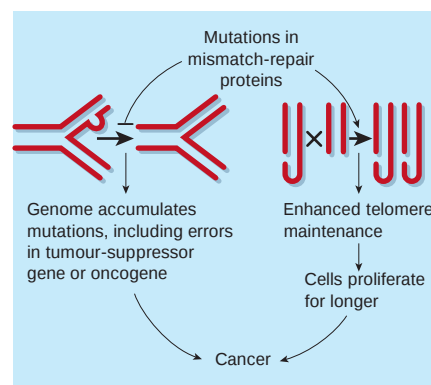


Figure 1 Mismatch repair in cancer. The mismatch-repair system has two functions: fixing errors in DNA, and blocking recombination between non-identical DNA sequences. As Rizki and Lundblad² show, one type of recombination that is blocked (at least in yeast) is that between the 'telomeric' DNA sequences that cap chromosome ends. So, defects in mismatch repair both prevent errors from being fixed (left) and allow telomeres to be replenished (right). Both of these mechanisms might contribute to human cancer.

Another is the increase in detrimental DNA rearrangements that occur as the result of recombination between non-identical (homeologous) DNA strands⁴.

A second pathway that maintains genomic stability works by replenishing the repeated DNA sequences, known as telomeres, that cap the ends of chromosomes. In most eukaryotes (organisms other than bacteria), this process relies on an enzyme called telomerase⁴. Most non-reproductive cells in humans do not express telomerase and, in the face of continuous cell proliferation, this deficiency leads to the progressive



100 YEARS AGO

Taxidermy; Comprising the Skinning, Stuffing and Mounting of Birds, Mammals and Fish. Edited by P. N. Hasluck. Certainly, the professional taxidermist, who has at his command works of the class of Mr. J. Rowley's *Art of Taxidermy* (reviewed in *Nature* for 1898), has nothing to learn from the present handbook, and it is difficult to imagine in what way the ordinary amateur is likely to be interested in the mounting of animals of the size of a waterbuck. It is not as if the author (or editor) had any new ideas to communicate with regard to the mounting of such mammals. On the contrary, although he confuses his readers with an unnecessarily complex system of measurements to be taken before skinning, he is really far behind advanced modern methods in his system, which bears no comparison with that adopted by many Continental and American taxidermists. Indeed, mediocrity may, in our opinion, be regarded as the leading feature of the book... As regards the skinning and stuffing of ordinary birds and the smaller mammals, the methods and descriptions are, in an old-fashioned way, well enough; and had the editor restricted himself to work of this nature not much fault could be found with his attempt. One thing we are glad to notice, namely, that the author advocates painting stuffed fish in imitation of their natural colours instead of being content with the faded scarecrows still to be seen some of our museums. Whether, however, the methods, both of mounting and colouring, advocated by him would result in the production of specimens bearing any real resemblance to their living prototypes could be decided only by actual inspection of the work.
From *Nature* 6 June 1901.

50 YEARS AGO

If each period in the periodic system ends in a rare gas (H, He 1st period, Li-Ne 2nd, etc.), then the number of elements in a period can be found from the ordinal number of the period by the following formula:

$$L = [(2n + 3 + (-1)^n)^2]/8$$

where n is the number of the period and L is the number of elements in the period. This formula also gives the relation between the number (n) of electron shells in an atom, and the number (L) of the elements in the period in which the particular atom is placed.
From *Nature* 9 June 1951.

erosion of telomeric repeats. Eventually, severe erosion leads to loss of the telomeres' capping function, resulting in cellular growth arrest and death, and rampant chromosomal instability. This instability propels 'aspiring' cancer cells towards the threshold of cancer-promoting changes^{5,6}. However, once over that threshold, some measure of chromosomal stability is ultimately required. So it is no surprise that most cells from advanced human tumours reactivate telomerase as a means of restabilizing chromosome ends, allowing such cells to survive and continue multiplying⁷.

Studies of yeast suggest that the reactivation of telomerase may not be the only means of replenishing diminished telomeres. In both budding and fission yeast, long-term cellular proliferation can occur even in the absence of telomerase, because recombination between chromosome ends provides another means of telomere maintenance, called alternative lengthening of telomeres^{8,9}. A similar mechanism for maintaining telomeres seems to work in telomerase-defective immortalized human cells¹⁰.

The mismatch-repair system and telomere-maintenance mechanisms have traditionally been thought to make independent contributions to genomic stability. But Rizki and Lundblad² reveal an intriguing relationship between these two processes, at least in yeast. In both yeast and human cells, the ends of chromosomes are not perfectly alike, suggesting that the alternative telomere-maintenance mechanism might involve homeologous recombination. So Rizki and Lundblad proposed that the mismatch-repair machinery inhibits recombination between non-identical telomeric sequences. They also suggested that removing this block to homeologous recombination might promote cell proliferation by allowing telomeres to be replenished in the absence of telomerase.

To test this model, the authors compared the proliferative potential of budding-yeast cells that lacked active telomerase, that had defects in the mismatch-repair machinery, or both. They show that the survival of yeast cells in the absence of telomerase was enhanced by defects in mismatch repair. The higher survival rate resulted from the removal of a block to homeologous recombination, rather than from the increased mutation rates that would also be seen in the absence of mismatch repair. As budding yeast and humans both have mismatch-repair proteins, and both have repeated telomeric sequences, it is likely that this result will apply to human telomeres. This seems even more plausible because several mutations in a yeast mismatch-repair gene that correspond to mutations seen in HNPCC patients also enhance the survival of normally ageing yeast cells².

These results have implications for under-

standing telomerase-independent telomere maintenance. Most yeast (and human) cells do not survive the loss of telomeres. So, in a culture of mutant yeast cells that lose telomeres rapidly, most cells die, and only those that have activated a recombination-dependent pathway for telomere maintenance survive⁸. The survivors display dramatic amplification and rearrangements of their telomeric and subtelomeric repeats. A burst of recombination occurs just as the yeast cells with lengthy telomeres appear, and it has been suggested that critically short telomeres are the main targets of these rescuing recombination events¹¹. But Rizki and Lundblad² show that the restorative effects of the loss of mismatch repair on a telomerase-deficient culture are seen soon after the loss of telomerase, when telomeres have not shortened dramatically.

The new work² also enhances our understanding¹ of how defects in the mismatch-repair system contribute to tumour development in mammals (Fig. 1). A deficiency in mismatch repair leads to a predisposition to cancer at several anatomical sites. In many of these cancers, the oncogenes or tumour-suppressor genes that might be mutated as a result of defects in repairing damaged DNA have not yet been identified. Rizki and Lundblad's results suggest that, at least in some of these cases, a loss of mismatch-repair genes might have increased the time in which cells that lack telomerase can proliferate, by helping to keep telomeres long. Because the cells survive for longer, it is more likely that they would accumulate tumour-promoting mutations. Mice with mutations in mismatch repair as well as in essential components of telomerase are available, so these and related problems can now be tackled. On a broader note, this work is a prime example of how studies of yeast continue to guide our understanding of complex human diseases.

Raju Kucherlapati is in the Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA.

Ronald A. DePinho is at the Dana Farber Cancer Institute, Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA.
e-mails: ron_depinho@dfci.harvard.edu
kucherla@aecom.yu.edu

1. Harfe, B. D. & Jinks-Robertson, S. *Mutat. Res.* **451**, 151–167 (2000).
2. Rizki, A. & Lundblad, V. *Nature* **411**, 713–716 (2001).
3. Fishel, R. & Kolodner, R. D. *Curr. Opin. Genet. Dev.* **5**, 382–396 (1995).
4. Blackburn, E. H. *Nature* **408**, 53–56 (2000).
5. Artandi, S. E. et al. *Nature* **406**, 641–645 (2000).
6. Rudolph, K. L., Millard, M., Rosenberg, M. W. & DePinho, R. A. *Nature Genet.* **28**, 155–159 (2001).
7. Shay, J. W. & Bacchetti, S. *Eur. J. Cancer* **33**, 787–791 (1997).
8. Lundblad, V. & Blackburn, E. H. *Cell* **73**, 347–360 (1993).
9. Nakamura, T. M., Cooper, J. P. & Cech, T. R. *Science* **282**, 493–496 (1998).
10. Dunham, M. A., Neumann, A. A., Fasching, C. L. & Reddel, R. R. *Nature Genet.* **26**, 447–450 (2000).
11. Tend, S.-C., Chang, J., McCowan, B. & Zakian, V. A. *Mol. Cell* **6**, 947–952 (2000).

Quantum electronics

Nanotubes go ballistic

Carter T. White and Tchavdar N. Todorov

As devices shrink, tiny wires that conduct electrons ballistically — without scattering — have exciting applications. Carbon nanotubes can now do this over hundreds, even thousands, of nanometres, with stunning results.

In the absence of other disturbances, any wave that hits two semireflective barriers, one after the other, will produce an interference pattern. This pattern consists of regular oscillations in the intensity of the transmitted wave across the double barrier, as a function of wavelength. On page 665 of this issue Liang *et al.*¹ report such oscillations in the transmission of electrons through a metallic single-walled carbon nanotube (SWNT) hundreds of nanometres long that is held between two electrodes.

This experiment demonstrates the quantum-mechanical wave nature of electrons. It also shows that the propagation of electrons in the nanotube is ballistic — largely free from scattering — over distances of thousands of atoms. A few years ago, we predicted theoretically the possibility of

ballistic propagation of electrons over such distances in metallic SWNTs². The creation by Liang *et al.* of a device that relies on this effect for its operation confirms this² and other results^{3–6} that indicate ballistic transport through metallic carbon nanotubes, and is a stunning achievement.

Figure 1a shows what happens when a wave, ψ , hits two successive barriers, B_1 and B_2 . A classical or quantum wave is described by its amplitude and phase. The amplitude sets the heights of the peaks and troughs, and the phase sets their positions at a given time. In Fig. 1a each transmitted partial wave, ψ_{T_n} , differs from the previous partial wave, $\psi_{T_{n-1}}$, by two extra reflections plus a round trip between the barriers. Each internal reflection reduces the amplitude of the wave by a fixed fractional amount and causes a fixed

change in phase. The round trip between the barriers adds a further phase change of $2\pi L/\lambda$, where L is the length of the round trip and λ is the wavelength of the wave between B_1 and B_2 . The rules of wave superposition then tell us that the intensity of the total transmitted wave, ψ_T , is a periodic oscillatory function of $1/\lambda$, with a period determined by L . If the wave is a beam of light hitting two partially reflecting mirrors, then the device is the basis of a Fabry–Perot interferometer⁷, which is widely used in astronomy and optics for accurately measuring wavelengths. Or ψ could be a sound wave hitting two walls, or a water wave passing over two hillocks on the seabed.

In its quantum-mechanical guise, ψ could be the wavefunction describing an electron in a ballistic quantum wire between two electrodes, as shown in Fig. 1b. Partial reflections at the contacts between the wire and the electrodes will then produce regular oscillations in the intensity of the total transmitted electron wave from one electrode to the other as a function of the wavelength of the electrons. A theoretical framework developed by Landauer and Büttiker⁸ to describe electrical conduction in quantum wires allows us to relate this transmitted intensity directly to the electrical current, I , flowing in such a system under an applied voltage, V , between the electrodes.

In their experiment, Liang *et al.* show that a metallic SWNT between two electrodes, with weakly reflecting contacts, exhibits regular smooth variations in the differential conductance G , where $G = dI/dV$, as a function of V and of the gate voltage, V_g . The applied voltage, V , controls the total energy with which electrons enter the nanotube, whereas V_g controls their potential energy in the nanotube. Therefore both V and V_g control the kinetic energy of the electrons that contribute to the current and hence also their wavelength. The wave interference analysis is now more complex because there are two possible quantum states for electron conduction through the nanotube⁹, but the principles remain the same.

Any significant scattering of the electrons within the nanotube would affect the amplitudes and phases of the partial electron waves transmitted into the far electrode, and so would affect the observed conductance interference pattern. The results of Liang *et al.* therefore provide evidence that electrons traverse the length of the nanotube — hundreds of nanometres, or thousands of atoms — ballistically, without significant scattering. Thus their device can be thought of as a resonant cavity for electrons in which the metallic SWNT acts as a waveguide and the contacts between the nanotube and the electrodes act as weakly reflecting barriers, as their model shows.

Devices such as the one created by Liang *et al.* are possible even if there is residual

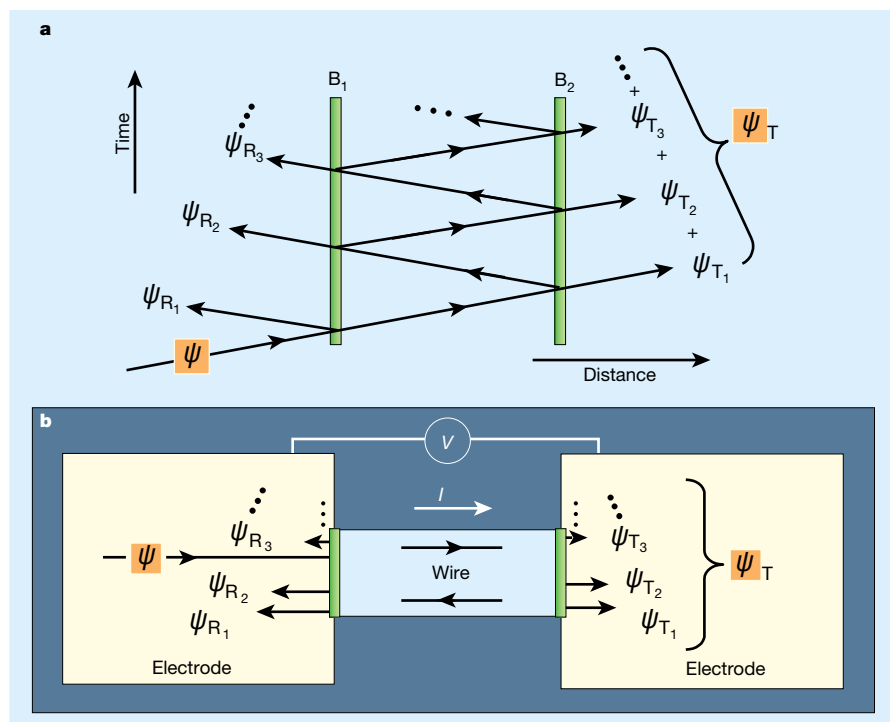


Figure 1 A wave hitting two partially reflective barriers. a, Multiple reflections of a wave, ψ , incident on two barriers, B_1 and B_2 . The wave arrives at B_1 and splits. Some of it is reflected straight back, forming the partial wave ψ_{R_1} . The rest reaches B_2 , where it is partially transmitted, forming ψ_{T_1} . The rest is reflected back to B_1 , where it is partially transmitted, forming ψ_{R_2} . The rest returns to B_2 , where it is partially transmitted, forming ψ_{T_2} , and so on. The total transmitted wave on the right of B_2 , ψ_T , is a superposition of all transmitted partial waves: $\psi_T = \psi_{T_1} + \psi_{T_2} + \dots$. In this series, ψ_{T_n} has undergone two extra internal reflections plus a round trip between the two barriers compared with $\psi_{T_{n-1}}$. b, Transmission of an electron wave, ψ , through a length of ballistic wire between two electrodes with partially reflecting contacts. The electron wave undergoes multiple reflections back and forth along the wire between the two wire–electrode contacts. In the experiment of Liang *et al.*¹, the ballistic wire is a metallic single-walled carbon nanotube hundreds of nanometres long.

structural and chemical disorder in the nanotube. This follows because electrons travelling in the nanotube do not experience this disorder, but rather its smoothed average over the tube's circumference. This implies that the characteristic distance over which ballistic transport can be sustained by the tube, l , will increase for tubes of larger diameter². There are, of course, some practical limits, such as the tendency of tubes with too large a diameter to collapse. But this relationship between l and diameter means that metallic tubes with nanometre diameters, such as those used by Liang *et al.*, are probably ballistic over distances exceeding a micrometre².

The increase in l as a result of better averaging of the disorder in larger-diameter tubes is possible only because of a remarkable property of metallic SWNTs: the number of quantum states available for electron conduction through the tube is independent of the tube's diameter². This key property allows conduction electrons in larger-diameter tubes to take advantage of better averaging over the tube's circumference without confronting additional states for electron backscattering. Such additional states would appear in normal metallic wires of increasing cross-section, eliminating the benefits of disorder averaging.

Semimetallic SWNTs (a family of SWNTs that are not quite metallic because of their intrinsic structure) can also share this property of metallic SWNTs, and so should show a similar relationship between l and diameter. This key property of metallic SWNTs is also intimately connected to their stability against a Peierls distortion (a spontaneous symmetry breaking that would turn them into semiconductors) and hence is central to their ability to conduct⁹. Additionally, their

special properties mean that conduction in metallic SWNTs should be insensitive to long-range disorder^{5,6} and bending^{2,10}.

With metallic SWNTs, nature has provided us with mechanically robust and flexible electron waveguides. The ballistic character of these nanometre-diameter wires suggests that they might be resistant to electromigration — atomic rearrangements and diffusion — caused by current-induced forces on atoms in regions of electron scattering¹¹. Electromigration damage is a central failure mechanism in conventional metallic wires used in electronic circuitry, and is among the harshest problems faced by the electronics industry. Although ballistic transport through the nanotube waveguide appears robust, it will be affected by the environment of the tube over long enough distances. Hence, devices such as that created by Liang *et al.* — especially those with greater separation between the electrodes — could be used as miniature strain gauges, or as detectors of trace amounts of foreign chemical species.

The apparent ability of metallic SWNTs to transmit quantum-mechanical electron waves without loss of information might also attract attention in the emerging field of quantum computing. ■

Carter T. White is in the Chemistry Division, Naval Research Laboratory, Washington DC 20375-5320, USA.

Tchavdar N. Todorov is in the School of Mathematics and Physics, Queen's University of Belfast, Belfast BT7 1NN, UK.

e-mails: carter.white@nrl.navy.mil

t.todorov@qub.ac.uk

1. Liang, W. *et al.* *Nature* **411**, 665–669 (2001).
2. White, C. T. & Todorov, T. N. *Nature* **393**, 240–242 (1998).
3. Tans, S. J. *et al.* *Nature* **386**, 474–476 (1997).
4. Frank, S., Poncharal, P., Wang, Z. L. & De Heer, W. A. *Science* **280**, 1744–1746 (1998).
5. Ando, T. & Nakanishi, T. *J. Phys. Soc. Jpn* **67**, 1704–1713 (1998).
6. McEuen, P. L. *et al.* *Phys. Rev. Lett.* **83**, 5098–5101 (1999).
7. Hecht, E. *Optics* (Addison-Wesley, Reading, MA, 1987).
8. Büttiker, M., Imry, Y., Landauer, R. & Pinhas, S. *Phys. Rev. B* **31**, 6207–6215 (1985).
9. Mintmire, J. W. *et al.* *Phys. Rev. Lett.* **68**, 631–634 (1992).
10. Kane, C. L. & Mele, E. J. *Phys. Rev. Lett.* **78**, 1932–1935 (1997).
11. Todorov, T. N., Hoekstra, J. & Sutton, A. P. *Phil. Mag. B* **80**, 421–455 (2000).

Statistical physics

Advances in aggregation

Andrew Zangwill

The theory of how atoms and molecules diffuse and collide to form clusters and droplets is incomplete. A new approach can predict the growth of thin films on a surface.

Ants live in colonies, antelopes travel in herds, and people join fan clubs. It's good to be a member of a group. The same is true for atoms and molecules. Given a chance, they will spontaneously aggregate into aerosols, colloids, gels, suspensions,

clusters and solids. The reason they do this is clear: to gain binding energy and hence become more stable. Precisely how they go about the business of joining together is less obvious. For this reason, a new theory reported in *Physical Review Letters* by Amar, Popescu and Family¹ is notable because it uses a relatively simple method to generate predictions of aggregation behaviour that agree quantitatively with both experiments and large-scale computer simulations.

Early in the twentieth century, the eminent Polish physicist Marian von Smoluchowski proposed a theory of aggregation² that uses rate equations to describe the microscopic processes of diffusion, collision and irreversible coalescence of multiparticle aggregates. The parameters of the equations — the rate constants — determine how quickly the various kinetic processes listed above take place. Once these are supplied, the theory predicts how many big and small aggregates are present at any moment in time — that is, how the cluster-size distribution changes over time.

The aggregation of water droplets in clouds³ is a typical case for which Smoluchowski theory has been used to predict droplet-size distributions. It is typical because there is only modest agreement between theory and experiment. Uncertainties arise

Astronomy

Endless clear skies

Astronomers have used many tricks to get round the problem of atmospheric distortion: finding remote mountain tops for their telescopes or, most expensively, lobbing telescopes into space. An alternative solution — adaptive optics — uses computer-controlled mirrors to compensate for the effects of the atmosphere in real time. As shown by this infrared image of a multiple star system, adaptive optics has now reached even higher resolution than the Hubble space telescope can manage.

This image of two stars in the GG Tauri system was taken



by Dan Potter at the Gemini North Telescope in Hawaii. Adaptive-optics systems, like the University of Hawaii's Hokupa'a (which means North star or steady star), use a fixed guide star as the first step towards compensating for

atmospheric turbulence. The image of GG Tauri was further enhanced using dual imaging polarimetry.

The disk of gas and dust surrounding these two stars appears to have a gap on the right-hand side. It is hard to explain a stable gap in such a rotating disk, but it could be a transient feature. Computer simulations of such binary systems show that two distinct spiral arms can develop, which may also explain the gap. There was a hint of a clearing in previous Hubble pictures, but the Hokupa'a image confirms that it is real. Sarah Tomlin

because the theory is approximate, the rate constants are not known with great accuracy, and the experimental data on the droplet-size distribution are obtained indirectly.

A much better system for comparing aggregation theory and experiment is the nucleation and growth of vapour-deposited material on the surface of a single crystal. In the earliest stages before a single layer develops, deposited atoms diffuse on the bare surface and aggregate into 'islands' that are one atom high. It is fairly straightforward to measure the island-size distribution (as a function of the total amount of deposited material) by using a scanning tunnelling microscope to image every island individually⁴. It is similarly straightforward to develop a computer simulation that tracks the motion of every deposited atom as it walks randomly across the surface and attaches irreversibly to an immobile island. Numerical simulations of this type⁵ quantitatively reproduce the experimental island-size distributions, albeit at considerable computational cost.

The achievement of Amar *et al.*¹ is to show that the computationally cheaper Smoluchowski approach can also reproduce the experimental island-size distributions. The new theory captures a significant but subtle property of the measured size distributions known as dynamical scaling. That is, the size distributions acquired at different times (corresponding to different surface coverage) collapse onto a single curve when plotted using suitably scaled variables. This means that the (normalized) probability of finding an island on the surface with a size equal to, say, half of the average island size stays the same despite the fact that the average island size changes with time.

The new calculations exploit the idea⁶ that one should track the evolution over time not only of the size of each island but also of the size of each island's 'capture zone'. To picture these capture zones, we must consider what happens to the pool of atoms that diffuse over the part of the surface not covered by islands in Fig. 1. The deposition process adds atoms to the pool everywhere. The islands subtract atoms from the pool at the island edges by collision and attachment. As a result, the concentration of migrating atoms falls to zero (or nearly so) right at the island edges. The dashed lines are the locus of points where the atom concentration reaches a local maximum. This network of lines defines a capture zone around every island. In the absence of random fluctuations, the diffusing atoms in a given zone flow towards their destination island so that the latter eventually captures every atom in its zone. As islands grow during aggregation and new islands nucleate, so must existing capture zones grow and new capture zones nucleate.

Amar *et al.*¹ derive a set of rate equations for the areas of the capture zones and then

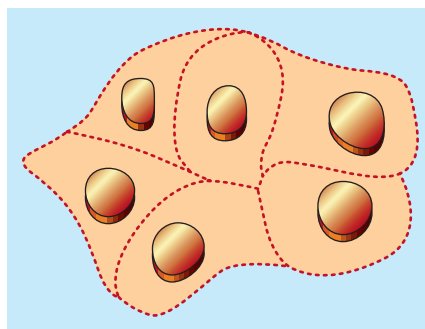


Figure 1 The deposition of a thin film onto a flat substrate is characterized by island formation. For the case when every atom–island collision results in the atom attaching irreversibly to an island edge, the local maxima of the atom concentration (dashed lines) define the boundaries of the island capture zones. Amar *et al.*¹ calculate the distribution of island sizes using a rate-equation theory that describes the evolution over time of the capture-zone areas as well as the evolution of the island sizes. (Adapted from ref. 9.)

combine them with the usual rate equations for the island sizes. To accomplish this, they calculate⁷ the rate constants for the collision processes, rather than input them as parameters. The high quality of the final results — particularly the dynamical scaling aspect — is intriguing because their rate-equation theory contains no randomness of any kind. In this respect, their work contrasts with recent computer simulations^{8,9} for the same problem, which suggest that dynamical scaling is somehow intimately connected with the random nature of the nucleation process. The issue is still open.

The agreement between experiment, simulation and rate-equation theory described by Amar and colleagues was achieved for the case when two colliding atoms are sufficient to form a stable island. In their study there are no thermal fluctuations that detach atoms from the edges of islands, and there is no diffusion of islands as a whole. It will be exciting to see if this latest version of the 80-year-old Smoluchowski theory can successfully model these more complex, and more realistic, aggregation processes. ■

Andrew Zangwill is in the School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. e-mail: andrew.zangwill@physics.gatech.edu

1. Amar, J. G., Popescu, M. N. & Family, F. *Phys. Rev. Lett.* **86**, 3092–3095 (2001).
2. von Smoluchowski, M. *Z. Phys. Chem.* **92**, 129–168 (1918).
3. Pruppacher, H. R. & Krell, J. D. *Microphysics of Clouds and Precipitation* (Reidel, Dordrecht, 1980).
4. Strosio, J. A. & Pierce, D. T. *Phys. Rev. B* **49**, 8522–8525 (1994).
5. Ratsch, C., Smilauer, P., Zangwill, A. & Vvedensky, D. D. *Surf. Sci.* **329**, L599–L604 (1995).
6. Mulheran, P. A. & Robbie, D. A. *Europhys. Lett.* **49**, 617–623 (2000).
7. Bales, G. S. & Chrzan, D. C. *Phys. Rev. B* **50**, 6057–6067 (1994).
8. Evans, J. W. & Bartelt, M. C. in *Morphological Organization in Epitaxial Growth and Removal* (eds Zhang, Z. & Lagally, M. G.) 50–72 (World Scientific, Singapore, 1998).
9. Ratsch, C., Gyure, M. F., Chen, S., Kang, M. & Vvedensky, D. D. *Phys. Rev. B* **61**, R10598–R10601 (2000).

Daedalus

Slow rubber

Among the important inventions of the modern world, a surprising number have been chemical. Daedalus is thinking of adhesive tape and paper, the sticky putty which has replaced them to some extent, and the 'rubber octopus' for fastening almost anything to almost anything else. He is now taking the obvious next step.

One problem these days is bringing fighter planes to a stop when they land on an aircraft-carrier. The solution is to use butyl rubber cables which engage with the incoming plane, absorb a lot of its energy, and contract fairly slowly. Scaling down this idea, DREADCO chemical engineers are making hollow butyl rubber tubes and filling them with Blutack or other adhesive putty compositions. The idea is to create a rubber that can swiftly be stretched, if need be, but that will then contract very slowly, ultimately exerting its full contractile force. With good fortune a suitably plasticized butyl rubber will combine both properties into a single composition.

Daedalus foresees a great many uses for a 'slow rubber' of this type. For a start, it will be ideal for rubber octopuses of all kinds. Instead of wrestling to get a suitcase on a roof-rack, for example, the motorist will merely need to stretch the DREADCO octopus to full size and lay it over the suitcase and the roof-rack. It will slowly contract until it grips the case as tightly as any conventional model. A small detachable winch, gripping the octopus, could stretch its arms, and allow extension forces that are hard to exert manually to be generated by simple mechanical means.

Many other uses will open up for the slow rubber. Cable-ties of all kinds will be made of it, allowing complicated wire connections to be insulated simply by slipping a stretched sleeve over the junction and waiting. Buttons and elastic inserts for pants and undergarments of all types could be fabricated from slow rubber, put on easily in the stretched form, and allowed to mould the figure at their leisure. Even the more ferocious type of corset could be made this way.

But of course the main use for slow rubber will be chosen by the market. Just as transparent plastic tape was initially directed to libraries for limited sale, so DREADCO's slow rubber will be launched in the hope that 'killer applications' will appear in due course. Daedalus himself hopes that the tyre-manufacturing industry will sit up and take notice.

David Jones

Successful invasion of a floral market

An exotic Asian plant has moved in on Europe's river-banks by bribing pollinators.

Invasive plants can displace native species through competition for nutrients, water, light and space. Here we show that they may also exert indirect pressure by competing for pollinators. We use the Asian plant *Impatiens glandulifera* as an example — this aggressive invader tempts bee pollinators away from native flowers with its rich nectar, which is more rewarding than that of any known native plant in central Europe. This causes a reduction of seed set in local plants in the vicinity and enables *I. glandulifera* to take over, reducing the fitness of native flora before competition for other resources even takes effect.

In just over a century since its introduction from the Himalayas, *I. glandulifera* (Fig. 1a) has conquered more than half of the river banks in the Czech Republic¹ and is spreading at a comparable rate elsewhere in Europe² and in North America³. At over 2 metres high, it is the tallest annual in Europe, with extensive branching of the main stem ensuring monopoly of the aerial environment. It also tolerates a wide range of climates and soil types⁴, which contributes to its success as a competitor. Each plant produces about 2,500 seeds, which fall to the ground at a density of 5,000–6,000 per square metre (ref. 2).

Pollination systems may be viewed as biological markets in which animals choose between 'products' (flower species) on the basis of their quality (sugar quantity in nectar, for example) and the plants compete for 'customers' (pollinators)^{4–6}. We tested whether competition for pollinators^{4,5} could be a mechanism by which *I. glandulifera* successfully excludes other plants. We first quantified the rewards they offer to pollinators. The nectar sugar concentration of *I. glandulifera* (48%) is well within the range of bumblebee-pollinated plants⁷, but the rate of sugar production (0.47 ± 0.12 mg per flower per hour; $n = 213$) is substantially higher than in the other common bumblebee-visited species that we found associated with *I. glandulifera* along river-banks near Würzburg. These include *Stachys palustris* (0.04 ± 0.02 mg per flower per hour; $n = 167$), *Lythrum salicaria* (0.02 ± 0.01 mg per flower per hour; $n = 85$) and *Epilobium hirsutum* (0.01 ± 0.02 mg per flower per hour; $n = 40$).

The rate of sugar production by *I. glandulifera* exceeds that recorded for any other central European plant — of the nectar-rewarding plant species, none produces more than of 0.3 mg per flower per hour and most generate less than 0.1 mg per flower per hour^{7–9}. *I. glandulifera* offers an

even richer reward than some flowers pollinated by hummingbirds (up to 0.27 mg per flower per hour by *Ipomopsis aggregata*¹⁰).

Stachys palustris occurs along river-banks in 'pure' patches, as well as in patches intermingled with *I. glandulifera*, a convenient arrangement for comparing the two species. Because of its extraordinarily rewarding nectar, visitation of *I. glandulifera* by bumblebees (mostly *Bombus pascuorum*) is frequent (2.5 ± 1.5 visits per flower per 10-min interval) — roughly fourfold more often than to *S. palustris* (Fig. 1b).

We simulated invasion by *I. glandulifera* by moving ten large *I. glandulifera* plants in water buckets into previously pure *S. palustris* patches (both plant types bore at least 30 flowers each). Bees readily accepted the new plants and, although visitation (0.88 ± 0.19 visits per flower per 10 min) was less frequent than in naturally occurring *I. glandulifera* patches, it was still greater than to *S. palustris* (Fig. 1b). We monitored the visitation frequency to *S. palustris* patches for 1–2 days before and 1–3 days after introducing *I. glandulifera* (excluding the first 3 h), by recording visitation to marked flowers in 7–50 short intervals (10 min per patch) according to a random schedule. We found that the number of visits to *S. palustris* was reduced by almost 50% when *I. glandulifera* was present (t -test, $t = 5.171$, $P = 0.0006$).

If the presence of *I. glandulifera* alters the visitation behaviour of pollinators to other plants, the fitness of these plants may also be affected. We therefore measured seed set in 102 inflorescences (with 2,567 flowers) from 15 pure patches of *S. palustris*, which were at least 500 m from the nearest *I. glandulifera* stand. We compared this seed set with that in 102 inflorescences (with 2,541 flowers) from 11 patches in which *S. palustris* grew intermixed with *I. glandulifera*. Seed set in the mixed patches was reduced by 25% relative to pure patches (Fig. 1c). A chi-squared contingency test on the frequency of flowers with 0, 1, 2, 3 or 4 seeds yielded a highly significant result ($\chi^2 = 21$; d.f. = 4; $P < 0.0001$).

To control for possible effects of competition for resources other than pollination, we repeated the experiment using potted *S. palustris* plants and found that, again, seed set in plants placed in *I. glandulifera* patches (mean, 2.29; s.d., ± 1.45) was significantly lower than in pure patches (mean, 3.2; s.d., ± 1.07 ; $\chi^2 = 16$; d.f. = 4; $P = 0.003$).

Our results show that the presence of *I. glandulifera* has a strong negative effect on the fitness of native plants, simply because it attracts many more pollinators. As such 'economic' choices may be made by bees

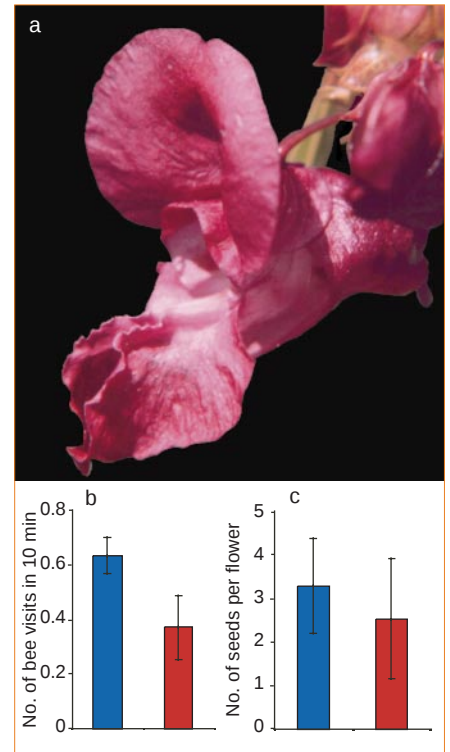


Figure 1 *Impatiens glandulifera*, an Asian beauty that has conquered Europe's river-banks. **a**, The strongly scented flowers of *I. glandulifera* are extremely rich in nectar and therefore very attractive to bumblebee pollinators. Photograph by J. Bitz. **b**, Pollinator visitation to *Stachys palustris* growing in uninterrupted patches (left) or in patches intermingled with *I. glandulifera* (right). **c**, Effect of the presence of *I. glandulifera* (right) on seed set of *S. palustris* (left, uninterrupted). *S. palustris* normally produces up to four seeds per flower. Error bars indicate standard deviation.

across large foraging ranges, competitive interactions between exotic and native plants may take place over much larger distances than previously thought. *I. glandulifera* and other exotic species may reduce the fitness of neighbouring plants even before competition for resources other than pollinators becomes an issue.

L. Chittka, S. Schürkens

Zoologie II, Biozentrum, Am Hubland,

D-97074 Würzburg, Germany

e-mail: chittka@biozentrum.uni-wuerzburg.de

- Pysek, P. & Prach, K. *Biol. Conserv.* **74**, 41–48 (1995).
- Beerling, D. J. & Perrins, J. M. *J. Ecol.* **81**, 367–382 (1993).
- Toney, J. C., Rice, P. M. & Forcella, F. *Northwest Sci.* **72**, 198–213 (1998).
- Waser, N. M. in *Handbook of Experimental Pollination Biology* (eds Jones, C. E. & Little, R. J.) 277–293 (Van Nostrand Reinhold, New York, 1983).
- Campbell, D. R. & Motten, A. F. *Ecology* **66**, 554–563 (1985).
- Schürkens, S. & Chittka, L. *Entomol. General.* **25**, 115–120 (2001).
- Comba, L., Corbet, S. A., Hunt, L. & Warren, B. *Ann. Bot.* **83**, 369–383 (1999).
- Comba, L. et al. *Ann. Bot.* **83**, 73–86 (1999).
- Corbet, S. A. et al. *Ann. Bot.* **87**, 219–232 (2001).
- Waser, N. M. *Ecology* **59**, 934–944 (1978).

Magnetic minerals

Keystone-like crystals in cells of hornet combs

The hexagonal brood-rearing cells inside the nest combs of the hornet *Vespa orientalis* are uniform in both their architecture and orientation. We have discovered that each cell contains a minute crystal that projects down from the centre of its domed roof and has a composition typical of the magnetic mineral ilmenite^{1,2}. These tiny crystals form a network that may act like a surveyor's spirit-level, helping the hornets to assess the symmetry and balance of the cells and the direction of gravity while they are building the comb.

The keystone-like crystal we describe here is glued with hornet saliva to the centre of the roof on the inside of each cell (Fig. 1); it is positioned away from the egg, which is fastened onto a side wall^{3,4}, the site at which the hatching larva also develops. The crystal is opaque, round, about 100 μm in diameter, and consists of polydomains. Diffraction analysis of the element composition (data not shown) reveals that these crystals contain large quantities of titanium,

iron and oxygen, with some carbon and trace amounts of manganese (58 out of 60 comb cells we inspected contained a crystal with the following composition in terms of number of atoms: 25–35% Ti, 15–20% Fe, 34–38% O, 12–18% C). This composition is similar to that of ilmenite crystals (FeTiO_3). The carbon component may originate from the hornet's saliva.

Comb building in social wasps occurs in stages. First, the hornet (whether a queen in the spring or workers thereafter) builds three to five concentric bands, which ultimately form the roof of the comb cell. Once the roof is completed, the hornet polishes it from the inside and deposits another construction layer of small particles or organic fibres. At the centre of the dome, the builder leaves a hollow where it attaches the crystal and which it fills with saliva that rapidly hardens into a polymer.

As the roof of the cell moves, the hanging crystals move in response to gravity. The downward-sloping cell walls that separate each domed roof from its six neighbours prevent the crystal from being shaken loose by workers or larvae communicating with other members of the colony, either acoustically or by tapping.

We presume that hornets collect these crystals from the local environment, but it is possible that they secrete them themselves, as titanium and iron are present in the hornet body⁵.

Ietse Stokroos*, Luba Litinetsky†, Johannes J. L. van der Want*, Jacob S. Ishay†

*Department of Cell Biology, Faculty of Medical Sciences, Graduate School of Brain Cognition and Neurosciences, University of Groningen, Oostersingel 69-2, 9713 EZ Groningen, The Netherlands

†Department of Physiology and Pharmacology, Sackler Faculty of Medicine, Tel Aviv University, Ramat-Aviv 69978, Israel

e-mail: physio7@post.tau.ac.il

1. Welham, N. J. *J. Mater. Sci.* **33**, 1795–1799 (1998).
2. Wilson, G. S. & Roberts, A. P. in *Paleomagnetism and Diagenesis in Sediments* (eds Tarling, D. H. & Turner, P.) **151**, 95–108 (Geol. Soc., London, 1999).
3. Ishay, J. S., Rosenzweig, E. & Paniry, V. A. *Insectes Sociaux* **29**, 34–43 (1982).
4. Spradbery, J. P. *Wasp* (Sidgwick & Jackson, London, 1973).
5. Ishay, J. S., Landsberg, A. & Pelah, S. *Scan. Microsc.* **10**, 187–208 (1996).
6. Jongebloed, W. L., Rosenzweig, E., Kalicharan, D., van der Want, J. J. L. & Ishay, J. S. *Electron Microsc.* **48**, 63–75 (1999).

Alzheimer's disease

Molecular consequences of presenilin-1 mutation

Alzheimer's disease is characterized by accumulation in the brain of a family of insoluble amyloid peptides ($\text{A}\beta$ peptides)¹, which are produced as a result of the normal processing of β -amyloid precursor protein (β -APP). Russo *et al.*² claim that a truncated $\text{A}\beta$ peptide that lacks the first ten amino acids accumulates in the brains of patients carrying a mutant form of presenilin 1 (PS1), a protein that is involved in cleavage of β -APP. However, we have found that this same species is also overrepresented in Alzheimer's patients with mutations in β -APP itself³. Our findings do not support the conclusion of Russo *et al.* that pathogenic PS1 mutations may control cleavage of β -APP by β -secretase^{4,5}.

Harmful mutations can occur at residue 717 in β -APP as well as in PS1, favouring generation of the $\text{A}\beta$ species $\text{A}\beta_{1-42}$ (refs 6, 7). Tandem missense mutations at residues 670 and 671 of β -APP promote cleavage at the $\text{A}\beta$ amino terminus⁸, the rate-limiting step in producing $\text{A}\beta$.

Russo *et al.*² characterized $\text{A}\beta$ species in the brains of patients suffering from sporadic or familial Alzheimer's disease due to mutations in PS1 or β -APP. Because N-terminally truncated $\text{A}\beta$ species (particularly those starting with glutamate at residue 11) were overrepresented in PS1-mutant brains of familial Alzheimer's sufferers, they inferred that pathogenic PS1 mutations affect β - as well as γ -secretase reactions.

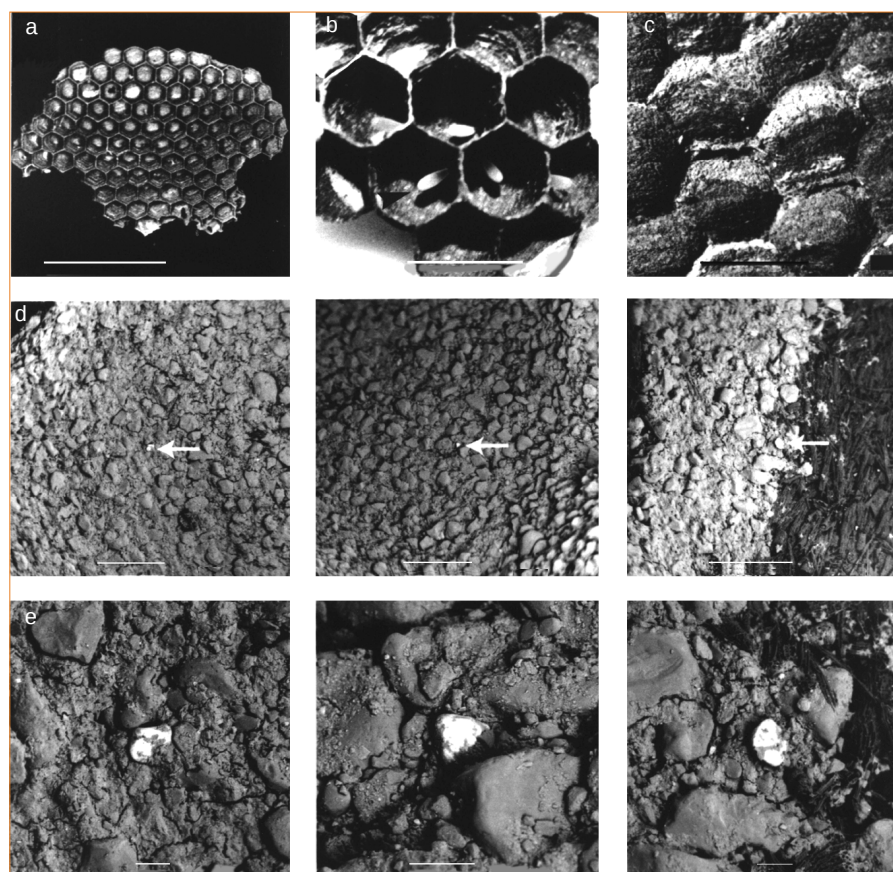


Figure 1 Macroscopic images and electron micrographs showing the locations of ilmenite crystals in cells of the comb of the oriental hornet. **a**, Inverted comb (with respect to the normal gravity-orientated position), showing the interior of comb cells and the inner side of the domed roof in each cell. **b**, Inverted hexagonal cells, some of which contain eggs. **c**, Small section of the comb at higher magnification, showing hexagonal roofs from the outside. **d, e**, At higher magnifications, an ilmenite crystal is seen (arrows in **d**) in the centre of the roof of different cells. For electron microscopy, the tops of cell walls were removed so that the roof could be seen; samples were prepared as described⁶. Scale bars: **a**, 5 cm; **b**, 1 cm; **d**, 1 mm; **e**, 100 μm .

We used size-exclusion chromatography and electrospray ionization mass spectrometry to study extractable amyloid in brains from elderly, non-demented individuals and from cases of either sporadic or familial Alzheimer's disease resulting from K670N/M671L or V717I mutations³ (where K, M and V represent the amino acid at that position in normal β -APP, and N, L and I the respective substituent in the mutant). We investigated their total A β content, whereas Russo *et al.*² looked primarily at soluble A β species.

Of the A β species we found, A β_{1-42} was the most abundant in four of five non-demented elderly brains, and A β_{1-40} was the most abundant in eight of ten brains with sporadic Alzheimer's. N-terminally truncated A β species were frequently identified in all patients: A β_{11-42} was the most abundant in one of ten sporadic patients and in the single β -APP(V717I) familial case, and the second most abundant in two of five non-demented individuals.

However, A β_{11-42} was more abundant than A β_{1-42} in four of ten sporadic patients and in those with V717I and K670N/M671L β -APP mutations. The K670N/M671L mutations at the A β amino terminus help β -secretase to cleave β -APP, yielding A β with aspartate as its first residue⁸, but in post-mortem tissue from a patient bearing these mutations, A β_{11-42} was the predominant species recovered from the deposited amyloid plaque, which is inconsistent with the conclusions of Russo *et al.*².

The Glu-11 A β species is the main peptide generated from endogenous β -APP in rodent neurons⁹ and in PS1-deficient mouse neurons¹⁰. The peptide bond at A β_{10-11} is also the preferred alternative cleavage site¹¹ for BACE, the protease that is responsible for most β -secretase reactions. These results argue against the model of Russo *et al.*² by showing that cleavage at this site occurs even in the absence of PS1.

N-terminally truncated A β peptides in amyloid plaques could be generated by post-production modification of the N terminus, rather than during cleavage by β -secretase. The pyrrolic modification of glutamate at residue 3 or 11 (ref. 12) of A β occurs after β -secretase cleavage. A β peptides that end at residue 42 may be relatively long-lived and so more susceptible to enzymes that cut at the N terminus, enhancing their heterogeneity by increasing their exposure to chemical modifications such as pyrolylation or oxidation.

Despite the technical difficulties in obtaining a full and accurate profile of all the A β species present in the amyloid plaques associated with various forms of Alzheimer's disease, we consider that amino-terminal truncation of A β is not unique to PS1-mutant familial Alzheimer's, and therefore question any proposed effect

of PS1 mutations on β -secretase activity.

Sam Gandy*, Jan Naslund†, Christer Nordstedt‡

*Sir James McCusker Alzheimer Research Unit, University of Western Australia, Perth, Australia and Department of Psychiatry, New York University School of Medicine, New York 10962, USA
e-mail: samgandy@earthlink.net

†Karolinska Institutet, 141 86 Huddinge, Sweden

‡AstraZeneca, 151 85 Södertälje, Sweden

1. Masters, C. L. *et al.* *Proc. Natl Acad. Sci. USA* **82**, 4245–4249 (1985).
2. Russo, C. *et al.* *Nature* **405**, 531–532 (2000).
3. Naslund, J. *et al.* *Proc. Natl Acad. Sci. USA* **91**, 8378–8382 (1994).
4. Li, Y. M. *et al.* *Nature* **405**, 689–694 (2000).
5. Esler, W. P. *et al.* *Nature Cell Biol.* **2**, 428–434 (2000).
6. Sherrington, R. *et al.* *Nature* **375**, 754–760 (1995).
7. Cai, X. D., Golde, T. E. & Younkin, S. G. *Science* **259**, 514–516 (1993).
8. Citron, M. *et al.* *Nature* **360**, 672–674 (1992).
9. Gouras, G. K. *et al.* *J. Neurochem.* **71**, 1920–1925 (1998).
10. Chen, F. *et al.* *J. Biol. Chem.* **275**, 36794–36802 (2000).
11. Vassar, R. *et al.* *Science* **286**, 735–741 (1999).
12. Miller, D. L. *et al.* *Arch. Biochem. Biophys.* **301**, 41–52 (1993).

Russo *et al.* reply — Gandy *et al.* compare our results¹ with their 1994 findings² that the amino-terminally truncated amyloid A β_{11-42} was relatively abundant in two cases of familial Alzheimer's disease involving two distinct mutations in β -APP. However, four important differences should be borne in mind: the authors compare A β_{11-42} with A β_{1-42} and ignore A β_{1-40} , although both A β_{1-40} and A β_{1-42} are generated by β -secretase/BACE cleavage at residue Asp 1 (ref. 3); their data are not correlated with features related to disease severity, such as age at onset and duration; they did not examine brains with PS1 mutations (these were not known at that time); and their characterization was based on the use of size-exclusion chromatography and electrospray mass spectrometry to quantify formic-acid-extracted A β , whereas we used quantitative analysis of immunoprecipitated water-soluble A β on western blots, mass spectrometry to identify A β variants, and immunohistochemistry to reveal amino-truncated A β peptides in plaques.

Table 1 gives ratios of the amounts of full-length A β_{1-40} and A β_{1-42} to that of A $\beta_{11-40/42}$ species in brains of Alzheimer's patients. The results of Naslund *et al.*² give values that are consistently greater than unity even when all other N-terminally truncated A β species they detected in sporadic Alzheimer brains are added to A β_{11-42} . The main difference with our data is in

sporadic rather than familial cases involving mutations in β -APP. The ratios between A $\beta_{1-40/42}$ and A $\beta_{11-40/42}$ that we found in PS1-mutant cases with early disease onset are much lower (less than unity) than those that Naslund *et al.* (and we) observed in familial Alzheimer's disease with β -APP mutations.

We proposed that increased formation of N-terminally truncated A β correlates with the early-onset phenotype of familial Alzheimer's disease linked to mutated PS1, and that a similar A β pattern may be shared by other malignant familial forms involving β -APP mutations^{4,5}. We explained the overrepresentation of these truncated A β forms in PS1-mutant brains by postulating that some PS1 mutations directly or indirectly affect β -secretase cleavage. A variety of mechanisms that do not require a direct effect of PS1 on β -secretase might be involved.

Gandy *et al.* suggest that late progressive proteolysis of the A β amino terminus occurs after β -secretase cleavage, after which glutamate is cyclized at the N-terminal residues 3 and 11. This sequence of events seems unlikely, as N-terminally truncated and pyroglutamylated carboxy-terminal fragments of β -APP (created before γ -secretase cleavage), which are direct precursors of similarly modified A β peptides, are present in the brains of Alzheimer's patients⁶. The finding that N-terminally truncated C-terminal fragments of β -APP are more abundant in PS1-deficient neurons⁷, quoted by Gandy *et al.* as evidence against our hypothesis, could also mean that PS1, either through its interaction with β -APP or its lack thereof, influences β -secretase cleavage. Some PS1 mutations may have the same effect on β -secretase cleavage as that produced by PS1 deficiency in a mouse knockout model.

We maintain that overrepresentation of N-terminally truncated A β forms may be characteristic of Alzheimer's disease that is associated with PS1 or other mutations and characterized by early onset and short duration, or other malignant phenotypic features. Whether PS1 mutations have an effect, be it direct or indirect, on β -secretase cleavage remains to be determined.

C. Russo, G. Schettini, T. C. Saido, C. Hulette, C. Lippa, L. Lannfelt, B. Ghetti, P. Gambetti*, M. Tabaton, J. K. Teller

Division of Neuropathology, Institute of Pathology,

Table 1 Ratios of A $\beta_{1-40/42}$ to A β_{11-42} in brains from Alzheimer's disease patients

Disease type	Naslund <i>et al.</i> ²	Russo <i>et al.</i> ¹
Sporadic	3.4 (n = 10)	1.4 (n = 17; 66–84 years)
APPK670M/N671L	1.7 (n = 1)	NA
APPV717I	1.2 (n = 1)	1.3 (n = 2; 59–63 years)
PS1 mutations (all)	NA	0.75 (n = 11; 40–81 years)
PS1 M139I and H163A	NA	0.3 (n = 4; 40–48 years)

Number of cases (n) and age at death are shown in parentheses. NA, not applicable.

Case Western Reserve University, Cleveland,
Ohio 44106, USA

*e-mail: pxg13@po.cwru.edu

1. Russo, C. *et al. Nature* **405**, 531–532 (2000).
2. Näslund, J. *et al. Proc. Natl Acad. Sci. USA* **91**, 8378–8382 (1994).
3. Vassar, R. *et al. Science* **286**, 735–741 (1999).
4. Ancolio, K. *et al. Proc. Natl Acad. Sci. USA* **96**, 4119–4124 (1999).
5. Crook, R. *et al. Nature Med.* **4**, 452–455 (1998).
6. Russo, C. *et al. Neurobiol. Dis.* **8**, 173–180 (2001).
7. Chen, F. *et al. J. Biol. Chem.* **275**, 36794–36802 (2000).

Nanoindentation

Simulation of defect nucleation in a crystal

Nanoindentation is the penetration of a surface to nanometre depths using an indenting device. It can be simulated using the Bragg bubble-raft model¹, in which a close-packed array of soap bubbles corresponds to the equilibrium positions of atoms in a crystalline solid. Here we show that homogeneous defect nucleation occurs within a crystal when its surface roughness is comparable to the radius of the indenter tip, and that the depth of the nucleation site below the surface is proportional to the half-width of the contact. Our results may explain the unusually high local stress required for defect nucleation in nano-indented face-centred cubic crystals.

Nanoindentation of face-centred cubic metals causes a load-versus-displacement response that is separated into regions of elastic deformation and discrete displacement bursts^{2–4} (Fig. 1a). The first displacement burst generally occurs when maximum shear stress generated under the indenter is of the order of the theoretical shear strength^{2,3}. This high local stress seems to cause homogeneous nucleation of dislocations beneath the surface,

producing a displacement burst³.

To confirm these observations, we used the bubble raft as a model for nanoscale atomic contact, in which the bubble positions represent the equilibrium positions of atoms¹ and allow visualization of deformation, dislocations, adhesion and fracture^{5–7}. We prepared the bubble raft as described^{1,6,7}. Indentation along the <121> direction of the raft proceeded orthogonally to the <110> closed-packed direction in the {111} plane. We flattened the raft's contact edge by removing extraneous bubbles with a soldering iron.

A completed raft, comprising more than 10⁴ bubbles, measures about 250 mm × 250 mm, and simulates semi-infinite boundary conditions. Each bubble was 1 mm in diameter, representing an atom of diameter 0.3 nm. All other relevant dimensions in the model are converted to atomic dimensions by using this conversion scale.

We indented a single-crystal bubble raft, which was initially defect-free, in the plane of the raft along the <121> direction by using indenters of simulated tip radii 8 and 28 nm. The indenter, which was constructed of aluminium plate, was positioned in the plane of the raft and slightly below the surface of the solution. We applied the load by controlled increase in displacement using a screw-driven mechanism (Fig. 1b). The in-plane shear stress beneath the indenter was maximal at a ratio of depth, z , to contact half-width, a , of 0.78 (Fig. 1c), as predicted by two-dimensional hertzian indentation theory⁸.

Figure 1c, d shows dislocation nucleation under atomically flat surfaces for two different indenter radii. In both cases, a dislocation nucleated beneath the indented surface along the loading axis at a depth of 0.78 a ; this location was determined by direct observation of z and a . The disloca-

tion then split in two: one dislocation glided into the crystal and the other ran to the surface, creating a slip step. Burgers circuits confirmed the dislocations as edge-type, with Burgers vector, $\mathbf{b}$, in the <110> direction. Dislocations nucleated at a constant depth of 0.78 a ; nucleation occurred farther below the surface as the indenter radius increased.

Our observations provide a valid assessment of stress distribution down to simulated atomic dimensions. This suggests that our approach could correctly describe elastic stress in actual nanoindentation, which is consistent with the elastic response seen before the first displacement burst during nanoindentation of face-centred cubic metals (Fig. 1a)^{2,3}.

Indentation of an atomically 'rough' surface initiates plasticity at the contact surface (Fig. 1e). Figure 1f shows nucleation inside the crystal, caused by indentation of a surface ledge of width comparable to the indenter radius. This trend supports the idea that roughness induces plasticity during nanoindentation of metals — that is, surface roughness (in the form of asperities and ledges) is expected to cause dislocation nucleation near the surface of the crystal if the width of the ledge is substantially smaller than the tip radius. If the width of the ledge is greater than the tip radius, available sites for heterogeneous nucleation are sufficiently distant from the point of maximum stress beneath the indenter to sustain dislocation nucleation within the crystal.

For typical nanoindentation with an indenter tip of radius 50 nm, surface asperities larger than 50 nm should show a similar indentation response to that for an atomically flat surface. As crystalline specimens are routinely polished to a roughness of 50 nm or more, this explains why the first displacement burst shown in Fig. 1 corresponds to stresses that approach the theoretical shear strength.

Andrew Gouldstone*, Krystyn J. Van Vliet, Subra Suresh

Department of Materials Science and Engineering,
Massachusetts Institute of Technology, Cambridge,
Massachusetts 02139, USA

e-mail: ssuresh@mit.edu

*Present address: Physiology Program,
Harvard School of Public Health, Boston,
Massachusetts 02215, USA

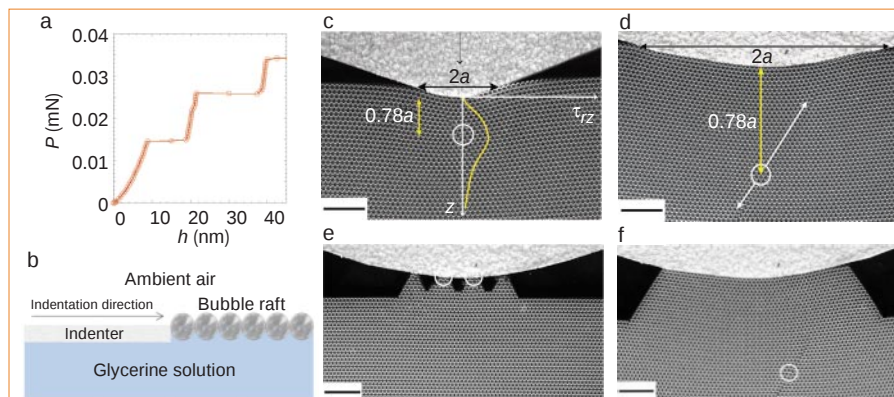


Figure 1 Dislocation nucleation during nanoindentation. **a**, Indentation load, P , plotted against depth, h , in (133) single-crystal aluminium, showing elastic deformation separated by displacement bursts attributed to dislocation nucleation³. **b**, Diagram of bubble-raft indentation. **c**, **d**, Initially defect-free rafts in which dislocation nucleation is induced by loading with indenters of different tip radii but at the same normalized depth, $z = 0.78a$ (white circle), where a is half the width of the contact. This is the position of maximum shear stress, τ_{xz} , shown in **c**. Nucleation (white circle) of a dislocation dipole is seen in **d**: one dislocation proceeds to the surface and the other into the crystal (white arrows). **e**, Indentation of atomic-scale asperities, showing plasticity (white circles) at the surface. **f**, Indentation of surface ledges of width comparable to the indenter radius causes dislocation nucleation within the crystal (white circle). Scale bar, 10 mm (representing 3 nm in our analogy).

1. Bragg, W. L. & Nye, J. F. *Proc. R. Soc. Lond. A* **190**, 474–481 (1947).
2. Suresh, S., Nieh, T. G. & Choi, B. W. *Scripta Mat.* **41**, 951–957 (1999).
3. Gouldstone, A., Koh, H. J., Zeng, K. Y., Giannakopoulos, A. E. & Suresh, S. *Acta Mat.* **48**, 2277–2295 (2000).
4. Gerberich, W. W., Nelson, J. C., Lilleodden, E. T., Anderson, P. & Wyrobek, J. T. *Acta Mat.* **44**, 3585–3598 (1996).
5. Shi, L. T. & Argon, A. S. *Phil. Mag. A* **19**, 255–274 (1982).
6. McClintock, F. A. & O'Day, W. R. *Proc. Int. Conf. Fract.* **1**, 75–98 (1965).
7. Georges, J. M., Meille, G., Loubet, J. L. & Tolen, A. M. *Nature* **320**, 342–344 (1986).
8. Johnson, K. L. *Contact Mechanics* (Cambridge Univ. Press, Cambridge, 1985).

Potassium channel receptor site for the inactivation gate and quaternary amine inhibitors

Ming Zhou, João H. Morais-Cabral*, Sabine Mann & Roderick MacKinnon

Howard Hughes Medical Institute, Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

Many voltage-dependent K^+ channels open when the membrane is depolarized and then rapidly close by a process called inactivation. Neurons use inactivating K^+ channels to modulate their firing frequency. In Shaker-type K^+ channels, the inactivation gate, which is responsible for the closing of the channel, is formed by the channel's cytoplasmic amino terminus. Here we show that the central cavity and inner pore of the K^+ channel form the receptor site for both the inactivation gate and small-molecule inhibitors. We propose that inactivation occurs by a sequential reaction in which the gate binds initially to the cytoplasmic channel surface and then enters the pore as an extended peptide. This mechanism accounts for the functional properties of K^+ channel inactivation and indicates that the cavity may be the site of action for certain drugs that alter cation channel function.

The presence of an inactivation gate causes a K^+ channel to close spontaneously after opening induced by membrane depolarization (Fig. 1a). The inactivation gate in Shaker-type K^+ channels is formed by the first 20 amino acids on the N terminus of the α -subunit^{1,2} or β -subunit³, located on the intracellular side of the membrane (Fig. 1c, d). The essential chemical characteristics that enable the N terminus to act as a gate are that the first approximately 10 amino acids tend to be hydrophobic (hydrophobic region) and the remaining 10 hydrophilic with excess positive charge (hydrophilic region)^{4,5} (Fig. 1d).

Four lines of evidence support the idea that the inactivation gate binds to the pore. First, inactivation occurs only after the voltage-dependent gate opens, as if the opening of the pore exposes a receptor for the gate⁶. Second, inactivation is produced by the binding of only one gate, presumably to the single pore opening, even though K^+ channels have four gates (N termini), by virtue of their homotetrameric architecture^{7,8}. Third, high concentrations of extracellular K^+ reduce inactivation, as if K^+ ions traversing the pore push the gate from its intracellular site⁹. Fourth, inactivation mimics the action of quaternary amines, which are thought to be pore blockers^{10,11} (Fig. 1a, b). Furthermore, quaternary amines compete with the gate to inhibit K^+ current¹¹.

How does the N-terminal gate interact with the pore to cause inactivation? Studies using mutagenesis have highlighted amino acids that might be expected to reside near the intracellular pore opening, for example, those between the fourth and fifth membrane-spanning segments, which connect the voltage sensor to the pore module¹². In addition, the structures of inactivation gates have been analysed by NMR spectroscopy^{13,14}. Together, these approaches have led to a picture of an inactivation 'domain' capping the pore's intracellular face^{13,15}. This picture is reasonable, but the quantitative details of earlier work are more compatible with a different structural view^{4,5}.

Pore occlusion by an extended N terminus

Structural studies have shown that the pore of a voltage-dependent K^+ channel opens to the cytoplasm between the T1 domain and the transmembrane channel^{16–18} (Fig. 1c). On the basis of mutant cycle

analysis, the inactivation gate was proposed to reach its site of action by entering the openings above the T1 domain¹⁶ (Fig. 1c). Here we ask, where is the inactivation gate's site of action? To address this question, we studied inactivation mediated by the $\beta 1$ subunit inactivation gate attached to the $\beta 2$ core ($\beta 12$)^{16,19} (Fig. 1c, d). The $\beta 12$ subunit was expressed in *Xenopus* oocytes with the $K_v1.4$ channel α -subunit, a mammalian homologue of the Shaker K^+ channel. The $K_v1.4$ α -subunit contained a deletion in its own N terminus (1.4-IR; see Methods) to ensure that inactivation would be mediated only by the $\beta 12$ inactivation gate^{16,19} (Fig. 1a). The inactivation process was parameterized by the inactivation time constants τ_{on} , τ_{off} and the ratio τ_{on}/τ_{off} , referred to as K_d (the dissociation constant; see Fig. 2a and Methods).

Mutations to alanine or to valine (position 6) in the inactivation gate affected K_d , a measure of the apparent affinity of the gate for its receptor, in a manner very consistent with previous findings on the Shaker K^+ channel^{1,2,4,5} (Fig. 2a). Mutations in the hydrophobic region had large energetic effects, expressed as changes in the apparent dissociation rate constant ($1/\tau_{off}$), whereas those in the hydrophilic region had more modest effects and altered both the apparent association ($1/\tau_{on}$) and dissociation rate constants. The importance of residues very close to the N terminus, in the hydrophobic region, is emphasized by the observation that a peptide corresponding to the first four amino acids alone (with a carboxy-terminal amide rather than a carboxylic acid) retains some ability to inhibit K^+ current (Fig. 2b).

The sixth membrane-spanning segment of voltage-dependent K^+ channels corresponds to the inner helix of the KcsA K^+ channel, which lines the pore on the intracellular side of the selectivity filter²⁰. This region of the pore forms a 10 Å-wide cavity at the membrane centre, the central cavity, that gradually tapers to about 4 Å diameter near the cytoplasmic opening. The pore-lining surface is predominantly hydrophobic in this region of the channel. We mutated amino acids that were predicted to point towards the pore on the basis of the KcsA K^+ channel structure. Five mutations out of six (to alanine) had significant effects on the inactivation gate K_d (Fig. 2c). The effects at positions 551 and 554, corresponding to cavity-lining residues 100 and 103 in KcsA, were so large that the K_d shown is an approximation based on the residual current measured after inactivation (see Methods). We used double-mutant cycle analysis of inactivation gate and inner helix mutations to assess the proximity of amino acids in the inactivated state (Fig. 2d). The Val 3 mutation

* Present address: Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, Connecticut 06520, USA.

on the gate was coupled to the Val 558 and Val 562 mutations on the inner helix, and the Ile 5 mutation was coupled to the Tyr 569 mutation. It was not possible to determine accurately the mutant cycle coupling energies involving positions 551 and 554, but the results at other positions imply that the inactivation gate lies in an extended conformation in the inner pore (Fig. 2d, e).

The central cavity binds hydrophobic cations

The above results support the simple conclusion that the inactivation gate apparently enters the inner pore and lodges its N terminus into the central cavity. To further support this hypothesis, we next made use of the fact that quaternary ammonium inhibitors mimic the action of the inactivation gate^{10,11} (Fig. 1a, b). The KcsA K⁺ channel was crystallized in the presence of tetrabutylammonium (TBA) and an electron-dense analogue, tetrabutylantimony (TBSb). TBSb is chemically very similar to TBA and blocks K⁺ channels accordingly (Fig. 3a). The heavy atom Sb provides the distinct advantage of easy identification in an electron density map. Data were collected from each crystal and a difference electron density map ($F_{TBSb} - F_{TBA}$)PHIcalc was calculated (Fig. 3b; see

Methods). The strong electron density peak reveals the binding site for TBA in the cavity. Refinement of the channel–TBA complex indicates that the presence of TBA in the cavity has little influence on the structure. Compared to the structure without TBA, the inner helices are drawn inwards towards the centre by a few tenths of an angstrom at the level of the cavity and are unchanged below the cavity (Protein Data Bank code 1J95).

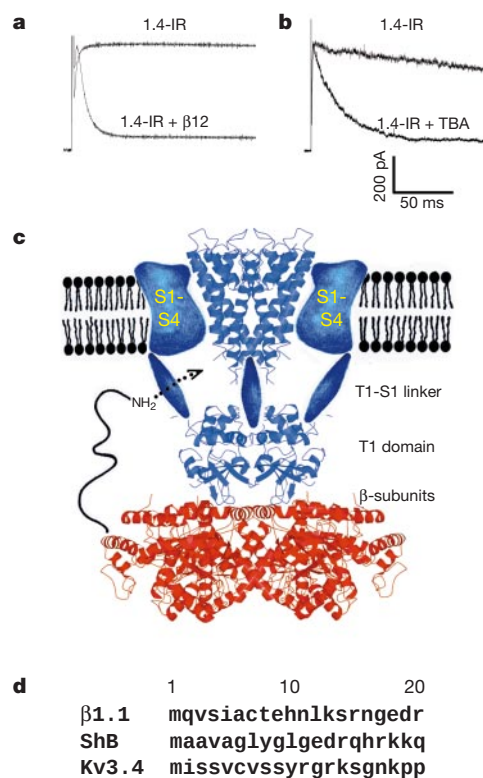


Figure 1 Biophysical features of K⁺ channel inactivation. **a**, K⁺ currents recorded from *Xenopus laevis* oocytes under two-electrode voltage clamp expressing channels without an inactivation gate (1.4-IR) or with inactivation gates provided by β -subunits (1.4-IR + $\beta 12$). The maximum current value is 1.4 μ A and 2 μ A for noninactivating and inactivating currents, respectively. Time scale is given in **b**. **b**, K⁺ currents from 1.4-IR channels recorded from an excised, inside-out patch under voltage clamp in the absence (1.4-IR) or presence of 10 μ M TBA (+ TBA). **c**, Composite model of a voltage-dependent K⁺ channel¹⁶. The α -subunit is shown in blue and the β -subunit in red. The pore is represented by the KcsA K⁺ channel²⁰ and the T1- β complex is from ref. 16. The structures of the voltage sensor (S1–S4) and linker (T1–S1) connecting the voltage sensors to the T1 domain are unknown. An N-terminal inactivation gate is shown entering a lateral opening to gain access to the pore. The image was prepared by Molscript³⁰ and raster-3D³¹. **d**, Sequence alignment shows inactivation gates from Kv1.1 (accession number CAA 50000), Shaker B (accession number CAA 29917) and Kv3.4 (accession number XP_002146).

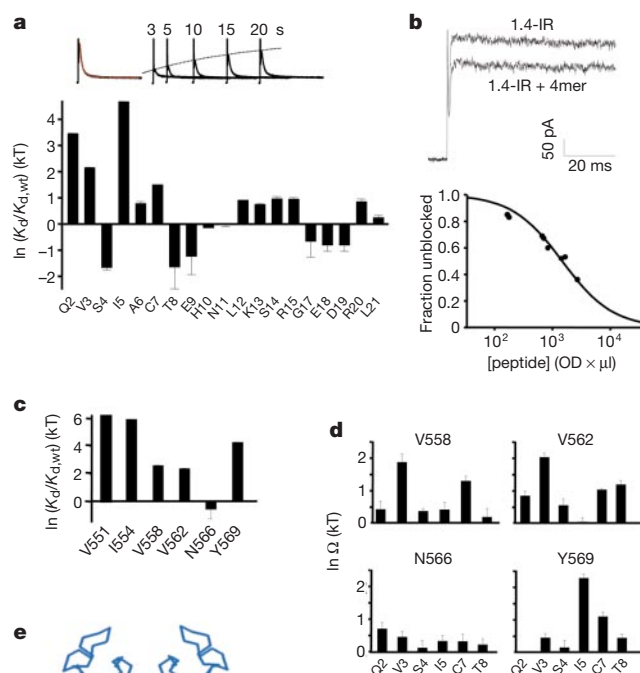


Figure 2 Mutational analysis of the inactivation gate–receptor interaction. **a**, Top, inactivation rates in Kv1.4-IR + $\beta 12$ channels determined by analysis of currents during a depolarizing pulse from –80 mV to +60 mV and recovery of current during a paired-pulse protocol⁷. τ_{on} (5.0 ± 0.3 ms) is the short time constant of a double exponential fit to current inactivation (red line) and τ_{off} (11 ± 0.7 s) is the time constant describing recovery in paired pulses (black line). Bottom, alanine-scanning mutagenesis of the inactivation gate. K_d , defined as τ_{on}/τ_{off} , was determined for Kv1.4-IR + $\beta 12$ channels with mutations to alanine or valine (position 6) at positions 2–21 in the $\beta 12$ inactivation gate. The K_d values, normalized by that for wild type, are shown. Error bars represent s.e.m. from ≥ 5 oocytes. **b**, Top, current recorded from an excised, inside-out patch containing 1.4-IR channels without (1.4-IR) and with (+ 4mer) a peptide corresponding to the first four amino acids of the $\beta 12$ inactivation gate. Bottom, dose–response curve showing current inhibition by the 4mer peptide as a function of concentration in units of optical density $\times$ volume. Data were collected from 12 patches. **c**, Alanine-scanning mutagenesis of pore-lining residues. The K_d for six pore-lining mutations to alanine, normalized by the wild-type K_d , is shown. Error bars represent s.e.m. from 3–7 oocytes. **d**, Double-mutant cycle analysis between pore-lining residues and residues on the inactivation gate. Ω calculated for six residues on the inactivation gate and four residues on the pore-lining helix is shown. Inactivation did not occur when Y569A on Kv1.4-IR was paired with Q2A on $\beta 12$. The approximate K_d determination for V551A and I554A mutations did not allow determination of Ω . Error bars show the s.e.m. measured in ≥ 5 oocytes. **e**, Summary of mutational analysis. Left, two diagonally positioned KcsA K⁺ channel subunits are shown in C α trace, with pore-lining residues of the KcsA K⁺ channel shown as sticks but labelled according to Kv1.4 residue numbering. Right, an extended strand model for the first six residues of the inactivation gate with side chains shown as sticks. Green and purple connecting lines identify coupled residues in the mutant cycle analysis.

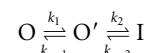
Inner helix mutations in the 1.4-IR channel alter inactivation and inhibition by TBA in a roughly similar manner (Fig. 3c). This finding further supports a common mechanism in which the inactivation peptide, like TBA, enters the pore and reaches the cavity. One notable difference between the inactivation gate and TBA is evident near the bottom of the inner helix, where the Y569A mutation has a relatively larger effect on inactivation. This difference is reasonable, however, as the inactivation gate comprises 20 amino acids and presumably interacts with the pore all the way from the cavity to the intracellular opening. A comparison of the size of TBA and the inactivation peptide leads us to propose that probably only the first three amino acids of the inactivation peptide bind in the cavity (Fig. 3d). These amino acids are generally hydrophobic in the inactivation gates (the $\beta 1$ inactivation gate is unusual in having a glutamine in the second position) and contain the N-terminal amino group, and are therefore chemically similar to TBA.

Our data lead us to propose that inactivation occurs through the interaction of the K^+ channel with a fully extended N-terminal peptide. The hydrophobic region of the peptide would extend from the cavity to the intracellular entryway, while the hydrophilic peptide region would emerge from the pore and interact with the aqueous protein surfaces lining the cage formed above the T1 domain (Fig. 1c). This configuration makes good chemical sense as the cavity and inner pore are lined by hydrophobic amino acids and the T1–S1 linkers outside the pore contain many acidic amino acids that would interact favourably with the inactivation peptide's multiple basic residues. This picture is consistent with the deduction of Aldrich *et al.* that the hydrophobic region must be 'buried

in a hydrophobic environment' and that the hydrophilic region is important for 'long-range electrostatic interactions'^{4,5}.

Sequential steps of inactivation

The idea that the N-terminal gate snakes into the inner pore as an extended peptide contradicts proposals of a structured inactivation domain docking superficially on the intracellular opening^{13,15}. The complex pathway to the intracellular pore opening indicates that the peptide probably reaches its final inactivating configuration through sequential binding steps (Figs 1c, 4a). For example, it is reasonable to imagine that the peptide first binds on the protein surface outside the pore, producing a preinactivated state, and then inserts its hydrophobic region to block the pore as outlined in the kinetic scheme (Fig. 4a):



The channel would conduct ions in the open (O) and preinactivated (O') states and become blocked in the inactivated (I) state. The effects of mutations on the forward (onset of inactivation) and backward (recovery from inactivation) rates can offer insight into the relative rates in such a sequential reaction scheme. Here and in previous studies, mutations in the hydrophobic region of the peptide affected mainly the apparent backward rate, $I \rightarrow O$, whereas mutations in the hydrophilic region affected modestly both the forward and backward rates^{4,5} (Fig. 4b). These observations make perfect sense if k_1 is small and $k_2 \gg k_{-1}$, conditions under which the forward rate $O \rightarrow I$ will be dominated by k_1 and the backward rate $I \rightarrow O$ will be related to $k_{-1} \times (k_2 / (k_2 + k_{-2}))$. Hydrophobic region mutations, by altering k_2 and k_{-2} , will affect only the backward rate, and hydrophilic region mutations, by altering k_1 and k_{-1} , will affect both rates. We therefore suggest that the first transition to form the

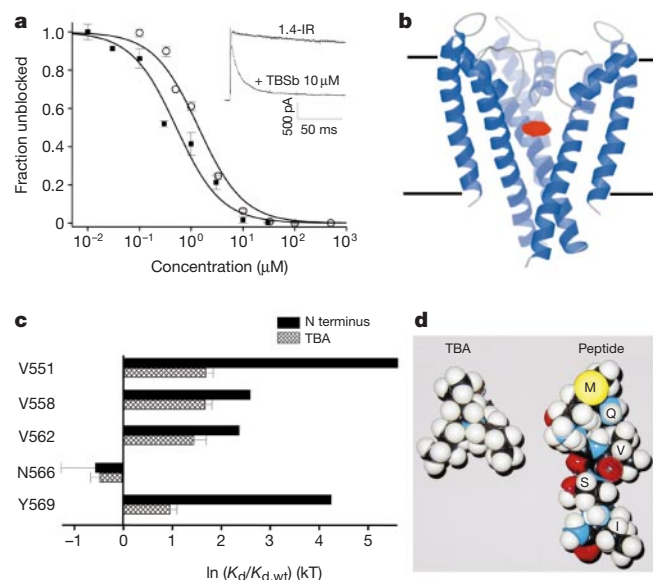


Figure 3 TBA binds in the cavity of the KcsA K^+ channel. **a**, Functional study of TBA and TBSb: currents from excised, inside-out patches were recorded under different TBA (open circles) or TBSb (filled squares) concentrations and fraction of residual current is plotted against inhibitor concentration. Smooth curves correspond to the Langmuir equation with K_d for TBA and TBSb of 1.5 μM and 0.7 μM , respectively. Inset, example currents with and without TBSb (10 μM). **b**, Ribbon diagram of three subunits of the KcsA K^+ channel. The red surface contoured at 6σ is the largest positive peak (maximum 18σ) present in a difference electron density map calculated at 4.0 Å with coefficients $F_{TBSb} - F_{TBA}$ and model phases from a refined model with TBA omitted. **c**, Effects of inner helix (S6) mutations on TBA block and inactivation. For pore-lining mutations on the inner helix, the natural logarithm of K_d for TBA and the inactivation gate (normalized to that of wild type) are plotted. Error bars for inactivation are s.e.m. from 3–7 oocytes. Error bars for TBA block are s.e.m. from ≥ 5 patches. **d**, CPK models of TBA (left) and the first five residues of the N-terminal inactivation gate (right). The inactivation gate is not at its most extended conformation, but the volume of the first three amino acids is comparable to that of TBA.

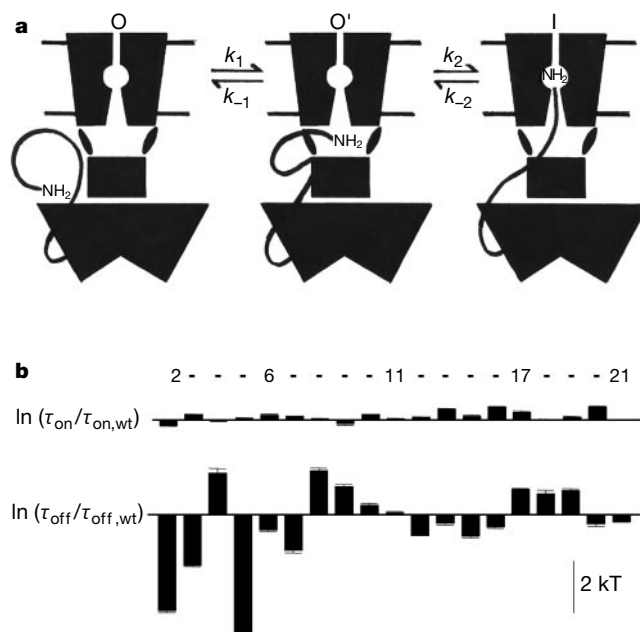


Figure 4 A structural model for the mechanism of inactivation. **a**, Open K^+ channel with three different configurations of the N-terminal inactivation gate shown attached to the β -subunit. For clarity, one inactivation gate is shown instead of four. O, open channel with its N terminus before docking; O', open channel with its N terminus bound to the hydrophilic protein surface; I, open channel with its N terminus entering the cavity (blocking the channel). **b**, The effect of N-terminal mutations on time constants of inactivation (τ_{on} and τ_{off}). The natural logarithms of inactivation time constants (normalized to that of the wild type) are plotted. Numbers above the bars represent inactivation gate residue numbers (Fig. 2). Error bars are s.e.m. from ≥ 5 oocytes.

preinactivated state can be rate limiting and that the final plugging transition can be fast, at least in some K^+ channels.

One prediction made by these rate conditions is that the forward rate of inactivation should be nearly voltage independent, as the rate-limiting step occurs outside the membrane electric field. Voltage-independent inactivation is observed in some K^+ channels⁶. A second prediction of these rate conditions is that if a fraction of channels exists in the preinactivated state before opening of the voltage-dependent gate, they should very rapidly inactivate upon opening. If sufficiently rapid, the inactivation peptide would appear to produce closed-state inactivation. Apparent closed-state inactivation has been documented with different inactivation gates of the Shaker splice variants⁴. Moreover, systematic truncation of the Shaker D peptide hydrophilic region modulated the extent of the apparent closed-state inactivation, as the model in Fig. 4a would predict⁴.

Discussion

Our findings lead us to conclude that the hydrophobic central cavity and inner pore of K^+ channels form the receptor site for both the inactivation gate and quaternary ammonium compounds. This conclusion explains the following functional properties of inactivation: a single gate is sufficient to cause inactivation^{7,8}; quaternary ammonium compounds compete with the inactivation gate¹¹; external K^+ pushes quaternary ammonium ions and the gate out of the pore^{9,21}; inactivation is voltage independent⁶; and some K^+ channels appear to exhibit closed-state inactivation⁴. A central cavity receptor for molecules as large as TBA and the inactivation gate also has implications for activation gating conformations in K^+ channels. In the KcsA crystal structure, the pore entryway near the cytoplasm has a diameter of about 4 Å. The diameter is unchanged when TBA is present in the cavity, but the pore must open sufficiently wide for TBA or the inactivation gate to reach the cavity.

Many pharmacological agents that influence cation channel function are both hydrophobic and cationic. On the basis of this study, we suggest that many of these agents bind in the cavity^{22,23}. □

Methods

Mutagenesis and expression

We used rat $K_v1.4$ -IR (residues 111–655, accession number CAA 34133) and rat $\beta 12$ chimera (rat $\beta 2$ core (residues 36–367, accession number CAA 54142) spliced at the N terminus with rat $\beta 1$ (residues 1–70, accession number CAA 50000))¹⁶. We introduced point mutations by the QuickChange method (Stratagene) and confirmed them by sequencing the entire complementary DNA insert. We prepared RNA by T7 polymerase transcription and injected it into *Xenopus laevis* oocytes²⁴.

Electrophysiology

We used a two-electrode voltage clamp (OC-725B, Warner Instrument Corp.) to record K^+ currents from oocytes 1–2 days after injection with messenger RNA. Electrodes had a resistance of ~ 0.5 M Ω (3 M KCl). The bath solution contained (in mM): NaCl 96, KCl 2, CaCl₂ 0.3, MgCl₂ 1 and HEPES 5 at pH 7.4. Oocytes were held at -80 mV and stepped to $+60$ mV for 200 ms to elicit K^+ current. Data collection and analysis methods are described in ref. 16.

We recorded patch-clamp currents from inside-out, excised patches from oocytes 3–5 days after injection. Electrodes were drawn from patch glass (PG150T-10, Warner Instrument Corp.) and polished to a resistance of 0.6–1 M Ω . The pipette solution (outside) contained (in mM): KCl 140, MgCl₂ 2 and HEPES 10 at pH 7.4. The bath solution (inside) contained (in mM): KCl 140, EGTA 5, MgCl₂ 2 and HEPES 10 at pH 7.4. K^+ currents were elicited by holding the patch at -100 mV and stepping to $+60$ mV for 300 ms. Solution exchange was achieved by gravity flow. Analogue data from an Axopatch-1D amplifier (Axon Instruments) were filtered (3 kHz, -3 dB) by an 8-pole Bessel filter (Frequency Devices), digitized at 20 kHz and stored on a PC hard disk.

Synthetic peptide and blockers

Inactivation peptide (4-mer) was synthesized by the Protein/DNA technology Center of the Rockefeller University. Rink amide resin was used to ensure that the C terminus of the peptide was amidated. We purified peptide by reversed-phase high-performance liquid chromatography, dissolved it in bath solution and added it directly to the bath. The peptide amount used was quantified by optical density (at 215 nm) $\times$ vol (μ l).

We purchased TBA and TBSb from Kodak and Aldrich, respectively. We dissolved TBA or TBSb in bath solution and perfused it onto an inside-out patch by gravity flow.

Crystallography

KcsA was expressed and purified as described²⁰. We incubated the chymotrypsin-cut protein at around 10 mg ml⁻¹ in a solution containing 150 mM KCl, 50 mM Tris (pH 7.5), 2 mM DTT and 5 mM *N,N*-dimethyldodecylamine-*N*-oxide with 1 mM of TBSb or 5 mM TBA for 15–30 min at room temperature. Crystals were obtained as described²⁰. Data were collected under a stream of boiled-off nitrogen at stations ID-13, ESRF and X-25, National Synchrotron Light Source, Brookhaven National Laboratory. The TBA co-crystal data extends to 2.8 Å with $R_{\text{sym}} = 7.2\%$, 93% complete, redundancy ~ 2 and the TBSb co-crystal data extends to 3.45 Å with $R_{\text{sym}} = 7.0\%$, 97% complete, redundancy ~ 3 . The data were processed with Denzo and Scalepack²⁵. All other calculations were done with the CCP4 package²⁶. The two data sets were scaled together to 4 Å with $R_{\text{merge}} = 22.5\%$ before the calculation of the difference electron-density map.

Analysis of inactivation, TBA block and double-mutant cycles

We determined the inactivation gate affinity for the channel by taking the ratio $\tau_{\text{on}}/\tau_{\text{off}}$. This definition approximates the equilibrium constant $k_{\text{off}}/k_{\text{on}}$ with two sources of error. First, even for a two-state process, $\tau_{\text{on}}/\tau_{\text{off}} = k_{\text{off}}/(k_{\text{on}} + k_{\text{off}})$. Given that in most channels studied $k_{\text{on}} \gg k_{\text{off}}$ this approximation introduces only a small error. Second, the inactivation process is not a two-state process, as discussed. In a separate analysis we modelled inactivation as a three-step reaction with a slow first and rapid second transition and found that our analysis, assuming two states, was sufficient for parameterization of mutational effects. To determine τ_{on} , we fit the inactivating current with a double exponential function and took the fast component ($\tau < 50$ ms, typically $> 90\%$ of inactivation) as τ_{on} . To determine τ_{off} , we fit the envelope of recovery in paired pulse experiments to a single exponential function⁷. In some mutant channels, we had to fit recovery with a double exponential function. We took the faster component ($\tau < 1$ s, 50–80% of current) as τ_{off} . Justification for this assignment is based on previous studies showing that the slow component of recovery is due to C-type inactivation^{19,27}. We verified this conclusion in two ways, by raising extracellular K^+ concentration to 96 mM, and by introducing a point mutation (K533Y) at the external TEA binding site^{9,28}. Both manoeuvres caused the slow component of recovery to disappear, compatible with its designation as the C-type process. In the case of mutations V551A and I554A, τ_{off} was too small to measure accurately and so we estimated the apparent gate affinity from the fraction of current remaining after inactivation ($\sim 87\%$ and $\sim 70\%$, respectively).

To quantify TBA, TBSb or peptide blocking, we plotted the fraction of residual current at the end of a 300-ms pulse against blocker concentration and fitted it with the following equation to obtain K_d , the equilibrium dissociation constant: fraction unblocked = $1/(1 + [\text{blocker}]/K_d)$.

We used the double-mutant cycle parameter Ω , where

$$\Omega = \frac{K_d^{\text{wt,wt}} \times K_d^{\text{mut,mut}}}{K_d^{\text{wt,mut}} \times K_d^{\text{mut,wt}}}$$

to quantify the degree of coupling between two mutants²⁹. An Ω value of more than unity indicates that the effects of two mutations are coupled. We used the mean and s.e.m. of K_d to obtain the range of uncertainty on Ω , assuming linear propagation of independent errors through the above equation.

Received 5 March; accepted 19 April 2001.

- Hoshi, T., Zagotta, W. N. & Aldrich, R. W. Biophysical and molecular mechanisms of Shaker potassium channel inactivation. *Science* **250**, 533–538 (1990).
- Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Restoration of inactivation in mutants of Shaker potassium channels by a peptide derived from ShB. *Science* **250**, 568–571 (1990).
- Rettig, J. *et al.* Inactivation properties of voltage-gated K^+ channels altered by presence of β -subunit. *Nature* **369**, 289–294 (1994).
- Murrell-Langado, R. D. & Aldrich, R. W. Interactions of amino terminal domains of Shaker K channels with a pore blocking site studied with synthetic peptides. *J. Gen. Physiol.* **102**, 949–975 (1993).
- Murrell-Langado, R. D. & Aldrich, R. W. Energetics of Shaker K channels block by inactivation peptides. *J. Gen. Physiol.* **102**, 977–1003 (1993).
- Zagotta, W. N. & Aldrich, R. W. Voltage-dependent gating of Shaker A-type potassium channel in *Drosophila* muscle. *J. Gen. Physiol.* **95**, 29–60 (1990).
- MacKinnon, R., Aldrich, R. W. & Lee, A. W. Functional stoichiometry of Shaker potassium channel inactivation. *Science* **262**, 757–759 (1993).
- Gomez-Lagunas, F. & Armstrong, C. M. Inactivation in ShakerB K^+ channels: a test for the number of inactivating particles on each channel. *Biophys. J.* **68**, 89–95 (1995).
- Demo, S. D. & Yellen, G. The inactivation gate of the Shaker K^+ channel behaves like an open-channel blocker. *Neuron* **7**, 743–753 (1991).
- Armstrong, C. M. Interaction of tetraethylammonium ion derivatives with the potassium channels of giant axons. *J. Gen. Physiol.* **58**, 413–437 (1971).
- Choi, K. L., Aldrich, R. W. & Yellen, G. Tetraethylammonium blockade distinguishes two inactivation mechanisms in voltage-activated K^+ channels. *Proc. Natl Acad. Sci. USA* **88**, 5092–2095 (1991).
- Isacoff, E. Y., Jan, Y. N. & Jan, L. Y. Putative receptor for the cytoplasmic inactivation gate in the Shaker K^+ channel. *Nature* **353**, 86–90 (1991).
- Antz, C. *et al.* NMR structure of inactivation gates from mammalian voltage-dependent potassium channels. *Nature* **385**, 272–275 (1997).
- Schott, M. K., Antz, C., Frank, R., Ruppersberg, J. P. & Kalbitzer, H. R. Structure of the inactivating gate from the Shaker voltage gated K^+ channel analyzed by NMR spectroscopy. *Eur. Biophys. J.* **27**, 99–104 (1998).
- Antz, C. & Fakler, B. Fast inactivation of voltage-gated K^+ channels: From cartoon to structure. *News Physiol. Sci. (Bethesda)* **13**, 177–182 (1998).
- Gulbis, J. M., Zhou, M., Mann, S. & MacKinnon, R. Structure of the cytoplasmic β subunit-T1 assembly of voltage-dependent K^+ channels. *Science* **289**, 123–127 (2000).

17. Kobertz, W. R., Williams, C. & Miller, C. Hanging gondola structure of the T1 domain in a voltage-gated K⁺ channel. *Biochemistry* **39**, 10347–10352 (2000).
18. Sokolova, O., Kolmakova-Partensky, L. & Grigorieff, N. Three-dimensional structure of a voltage-gated potassium channel at 2.5 nm resolution. *Structure* **9**, 215–220 (2001).
19. Heinemann, S. H., Rettig, J., Graack, H. R. & Pongs, O. Functional characterization of Kv channel beta-subunits from rat brain. *J. Physiol. (Lond.)* **493**, 625–633 (1996).
20. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
21. Armstrong, C. M. & Hille, B. The inner quaternary ammonium ion receptor in potassium channels of the node of Ranvier. *J. Gen. Physiol.* **59**, 388–400 (1972).
22. Yarov-Yarovoy, V. *et al.* Molecular determinants of voltage-dependent gating and binding of pore-blocking drugs in transmembrane segment IIIS6 of the Na⁺ channel alpha subunit. *J. Biol. Chem.* **276**, 20–27 (2001).
23. Mitcheson, J. S., Chen, J., Lin, M., Culbertson, C. & Sanguinetti, M. C. A structural basis for drug-induced long QT syndrome. *Proc. Natl Acad. Sci. USA* **97**, 12329–12333 (2000).
24. Lu, Z. & MacKinnon, R. A conductance maximum observed in an inward-rectifier potassium channel. *J. Gen. Physiol.* **104**, 477–486 (1994).
25. Otwinowski, Z. in *Proceedings of the CCP4 Study Weekend* (SERC Daresbury Laboratory, Daresbury, UK, 1993).
26. Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
27. Rasmuson, R. L. *et al.* C-type inactivation controls recovery in a fast inactivation cardiac K⁺ channel (Kv1.4) expressed in *Xenopus* oocytes. *J. Physiol. (Lond.)* **489**, 709–721 (1995).
28. MacKinnon, R. & Yellen, G. Mutations affecting TEA blockade and ion permeation in voltage-activated K⁺ channels. *Science* **250**, 276–279 (1990).
29. Hidalgo, P. & MacKinnon, R. Revealing the architecture of a K⁺ channel pore through mutant cycles with a peptide inhibitor. *Science* **268**, 307–310 (1995).
30. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
31. Merritt, E. A. & Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).

Acknowledgements

We acknowledge the European Synchrotron Radiation Facility (ESRF) and the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory (with support by the U.S. D.O.E., Division of Material Sciences and Division of Chemical Sciences). We thank C. Petosa and A. Perrakis for help on ESRF ID-13, and M. Becker for help on NSLS X-25. The project was supported by an NIH grant to R.M. R.M. is an investigator in the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to R.M. (e-mail: mackinn@rockvax.rockefeller.edu). Coordinates have been deposited with the Protein Data Bank under accession code 1J95.

An efficient photoelectric X-ray polarimeter for the study of black holes and neutron stars

Enrico Costa*, Paolo Soffitta*, Ronaldo Bellazzini†, Alessandro Brez†, Nicholas Lumb† & Gloria Spandre†

* Istituto di Astrofisica Spaziale del CNR, Via Fosso del Cavaliere 100, I-00133, Rome, Italy

† Istituto Nazionale di Fisica Nucleare-Sezione di Pisa, Via Livornese 1291, I-56010 San Piero a Grado, Pisa, Italy

The study of astronomical objects using electromagnetic radiation involves four basic observational approaches: imaging, spectroscopy, photometry (accurate counting of the photons received) and polarimetry (measurement of the polarizations of the observed photons). In contrast to observations at other wavelengths, a lack of sensitivity has prevented X-ray astronomy from making use of polarimetry. Yet such a technique could provide a direct picture of the state of matter in extreme magnetic and gravitational fields^{1–6}, and has the potential to resolve the internal structures of compact sources that would otherwise remain inaccessible, even to X-ray interferometry⁷. In binary pulsars, for example, we could directly ‘see’ the rotation of the magnetic field and determine if the emission is in the form of a ‘fan’ or a ‘pencil’ beam^{1,8}. Also, observation of the characteristic twisting of the polarization angle in other compact sources would reveal the presence of a black hole^{9–12}. Here we report the development of an

instrument that makes X-ray polarimetry possible. The factor of 100 improvement in sensitivity that we have achieved will allow direct exploration of the most dramatic objects of the X-ray sky.

The main advantage of the proposed polarimeter is its capability of investigating active galactic nuclei (quasars, blazars and Seyfert galaxies) for which polarization measurements have been suggested, crucial to understand the geometry and physics of emitting regions. We can separate synchrotron X-rays from jets^{13,14} from the emission scattered by the disk corona or by a thick torus. The effects of relativistic motions and of the gravitational field of a central black hole have probably been detected by iron line spectroscopy on the Seyfert-1 galaxy MCG-6-30-15 (ref. 15) but this feature is not ubiquitous in active galactic nuclei. Polarimetry of the X-ray continuum provides a more general tool to explore the structure of emitting regions^{16,17}, to track instabilities and to derive direct information on mass and angular momentum¹² of supermassive black holes.

In spite of this wealth of expectations, the important but only positive result until now is the measurement, by the Bragg technique, of the polarization of the Crab nebula^{18,19}. The Stellar X-ray Polarimeter²⁰ (SXP) represents the state of the art for conventional methods based on Bragg diffraction and Thomson scattering. However, Bragg polarimetry²¹ is dispersive (one angle at one time) and very narrow-band. Thomson polarimetry²² is non-imaging and band-limited (>5 keV). This limits the sensitivity of SXP to a few bright, galactic sources only.

The photoelectric effect is very sensitive to polarization. The electron is ejected from an inner shell with a kinetic energy which is the difference between the photon energy and the binding energy. The direction of emission is not uniform but is peaked around that of the electric field of the photons (see Fig. 1a). This photoelectron

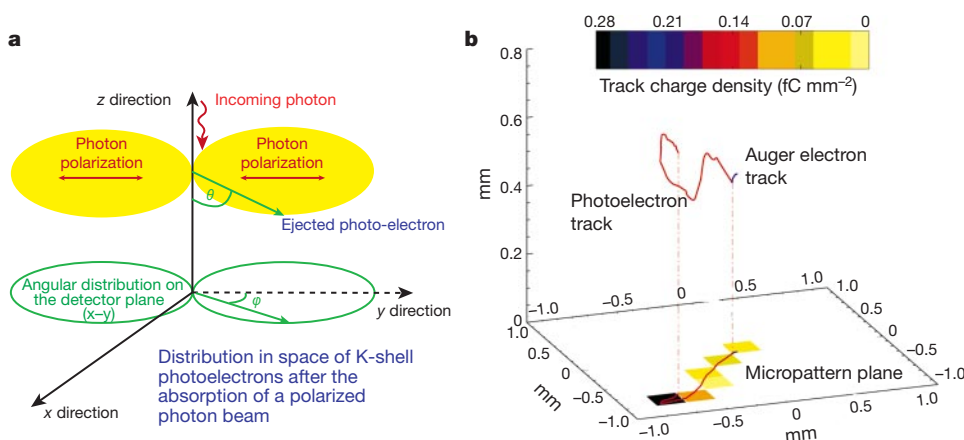


Figure 1 Basic physics of the photoelectric effect in a gas. **a**, Following the photon conversion in the gas, the photoelectron is ejected in directions that carry a significant memory of the electric field of the photon. When the beam is linearly polarized the electrons are ejected preferentially around the electric field. The cross-section of s electrons is:

$$\frac{\partial \sigma}{\partial \Omega} = r_0^2 \frac{Z^5}{137^4} \left(\frac{mc^2}{h\nu} \right)^{7/2} \frac{4\sqrt{2}\sin^2(\theta)\cos^2(\varphi)}{(1 - \beta\cos(\theta))^4}$$

where r_0 is the classical electron radius, Z is the atomic number of the target material and β is the electron velocity, as a fraction of the speed of light c . The figure shows the emission angle θ and azimuth angle φ . **b**, A simulated 6-keV photoelectron and Auger track, showing propagation in the gas and collection on a plane with sensitivity to both energy and position. The photoelectron is slowed by ionizing collisions with outer electrons of the atoms of the medium. The energy loss increases with decreasing kinetic energy (Bethe law for low energy).

$$\frac{\partial E}{\partial x} \propto \frac{1}{\beta^2} \propto \frac{1}{E_{\text{kin}}}$$

Electrons are also scattered by charges in the nuclei with no significant energy loss. This follows the screened Rutherford law:

$$\frac{\partial \sigma}{\partial \Omega} \propto \frac{Z^2/E_{\text{kin}}^2}{(\sin^2(\theta/2) + \alpha_{\text{screen}})^2}$$

Whereas scattering crucially depends on the atomic number, slowing down is only moderately dependent. The primary ionizations are then projected onto the sense plane. The charge density in each pixel is proportional to the energy loss, and is therefore related to the electron kinetic energy. The first part of the photoelectron path has a lower charge density, but it is closer to the initial direction of the photoelectron and so also closely related to the photon polarization direction. The second part has a higher charge density but it is randomized. The Auger electron track does not bring information on polarization.

interacts with the surrounding matter: it is slowed by ionizing collisions with atomic electrons and scattered by Coulomb diffusion on the nuclei (see Fig. 1b) and eventually stopped. The photoelectron leaves a string of electron-ion pairs in the absorber, marking the path from its creation to the stopping point. We call this string a 'track': in the initial part of this track resides the information on the original electron direction from which the polarization of the photon may be derived. This dependence is preserved if the track is projected onto a plane perpendicular to the radiation.

In a subdivided detector extended tracks may produce coincident signals in two contiguous cells. If the radiation is polarized the orientation of these pairs is asymmetric. We can exploit this by counting coincidences in neighbouring wires of proportional counters²³ or charge-coupled device (CCD) pixels^{24,25}. But as the detector cell is typically much larger than the electron range, the asymmetry effect strongly depends on the absorption point. This can be avoided if the cell is so small that the track is split into several cells. The first finely subdivided, self-triggered device²⁶ was a micro-gap²⁷ chamber filled with a neon-based gas mixture at 1 atm. The need to rotate the instrument and an as yet moderate modulation factor mean that this device is a step forward, but not by much.

Other instruments image the bright track made in a gas scintillation detector on a CCD. One of the two practical implementations²⁸ works only above 40 keV, the other²⁹ only at low pressure. Both use argon which, because of the higher energy of the isotropic Auger electron and larger multiple scattering, is a gas suitable for photon energies higher than the practical range for X-ray optics.

Position-sensitive gas detectors typically yield the centroid of the charge cloud, whose extent is the ultimate limit to the space resolution, a sort of noise to be kept as small as possible. We reverse this approach, trying to resolve the track to measure the interaction point and the prime direction of the photoelectron. To this end we have developed the micro-pattern gas chamber (MPGC). It consists (Fig. 2, Table 1) of a gas cell with a thick detection/drift region, a thin gas electron multiplier (GEM³⁰) and a multi pixel, true two-dimensional, read-out anode. The large number of fired pixels per track allows for a good track reconstruction. Additionally, the MPGC measures the energy lost in each pixel, a quantity directly related to the kinetic energy of the electron.

We filled the MPGC with a 1-atm mixture of Ne (80%)-dimethylether (DME; 20%). Figure 3a shows the image of a real MPGC track. An initial straighter part, with low ionization density, which carries

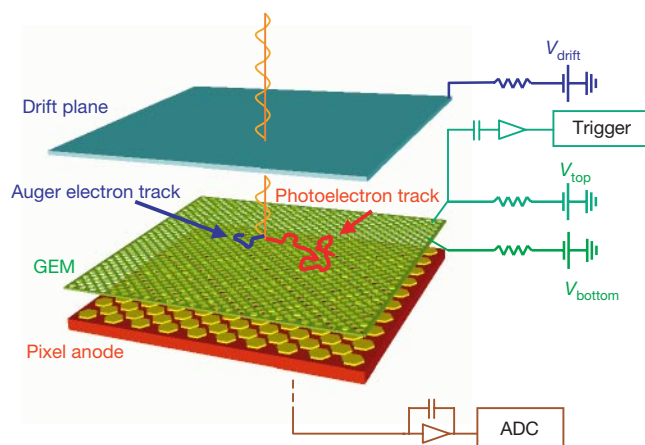


Figure 2 The micro-pattern gas detector. The photon is absorbed at some point in the drift gap. The photoelectron track is drifted by the electric field to the gas electron multiplier (GEM). This device is made of a thin (50 μm) polyimide foil perforated by many microscopic holes, where a high electric field provides the charge amplification. Finally, the charge is collected by the pixels of MPGC anode, each one connected to an independent electronics chain. On receiving a trigger from the GEM, all the signals are converted from analog to digital, so that we have the image of the track projected on the detector plane.

most of the information on the starting direction (and thence on the polarization) evolves into a skein with high ionization density and a completely random path. Most of the photoelectron energy is lost at some distance from the initial interaction point. To verify this interpretation we let impinge on the detector, through a very thin diaphragm, photons of 5.4 keV from an unpolarized source. The loci of the centroids of each track are displaced from the interaction point and located on a circular region around it, indicating that the tracks have, and retain, a significant elongation and energy loss asymmetry (Fig. 3b). From each track we reconstructed the emission angle and we built a histogram. In the case of unpolarized X-ray photons of 5.4 and 5.9 keV, all the emission angles have the same probability and the histogram is flat (Fig. 3b). When we irradiated the detector with an extended, nearly 100% polarized source of 5.4 keV, we found a strong angular modulation (44%) that is well modelled by the expected distribution (Fig. 3c), taking into account

Table 1 Physical characteristics and performances of the micro-pattern detector

	Present prototype (2–10 keV)	Improved configuration (3.5–10 keV)
Drift/absorption gap	6 mm	30 mm
Drift field	3,000 V cm ⁻¹	1,500 V cm ⁻¹
Gas filling and pressure	(Ne 80%–DME 20%); 1 atm	(Ne 40%–DME 60%); 4 atm
Gas grain	5,000	2,500
Transverse diffusion in drift	80 μm	<100 μm
GEM thickness	50- μm copper-clad kapton foil	50- μm copper-clad kapton foil
GEM hole geometry	40- μm diameter; 60- μm pitch	40- μm diameter; 60- μm pitch
GEM voltage	400 V	600 V
Detection efficiency at 5.4 keV	3.8%	91%
Read-out pixel size	200 μm	50 μm
Number of pixels	512	40,000
Read-out plane technology	Multilayer advanced PCB	VLSI
Track length/pixel size (6 keV)	6	6
Sensitivity to Her X1	$T = 400$ s; MDP = 10%	$T = 20$ s; MDP = 10%
Sensitivity to 3C-273	$T = 2.2 \times 10^5$ s; MDP = 2%	$T = 4 \times 10^4$ s; MDP = 1%
Sensitivity to MCG-6-30-15	$T = 5 \times 10^5$ s; MDP = 2%	$T = 1 \times 10^5$ s; MDP = 1%
Gain in the integration time over SXP (strong sources)	5 (over Thomson); 15 (over Bragg)	100 (over Thomson); 300 (over Bragg)
Gain in the integration time over SXP (faint sources)	250 (over Thomson); 100 (over Bragg)	5,000 (over Thomson); 2,000 (over Bragg)

We show also the observing time needed to measure (at 99% confidence) the shown degree of polarization (minimum detectable polarization or MDP) of a selected sample of astrophysical sources if the micro-pattern gas chamber (MPGC) is placed at the focus of the SODART³¹ telescope on board the Spectrum X-gamma mission. The relevant formalism can be found in ref. 26. For the improved design we use only photons above 3.5 keV to compute the sensitivity. Below this energy, at high pressure, the transverse diffusion and the unfavourable track length to pixel size ratio significantly reduces the modulation. Her X-1 is a galactic binary X-ray source with a magnetic field of about 3.5×10^{12} G measured by the detection of a cyclotron line. A high degree of linear polarization is expected and, by polarimetry, a direct measure of the angle between the magnetic field and the rotation axis with 1° accuracy can be performed. 3C-273 is the X-ray brightest radio-loud quasar and MCG-6-30-15 is a Seyfert-1 galaxy for which a broad ($\sim 80,000$ km s⁻¹) iron line was observed, possibly skewed by gravitational effects³². Finally, we compare the performances of MPGC with SXP, the most sensitive polarimeter foreseen in a future X-ray mission, ready to be installed in the focal plane of the SODART telescope. It simultaneously exploits the Bragg diffraction at 45° and the azimuth dependence on polarization of Thomson scattering. The ensemble analyser/detector of the two stages is rotated to search for a modulation in the counting rate and to compensate for systematic effects. GEM, gas electron multiplier. VLSI, very large scale integration; PCB, printed circuit board; DME, dimethylether.

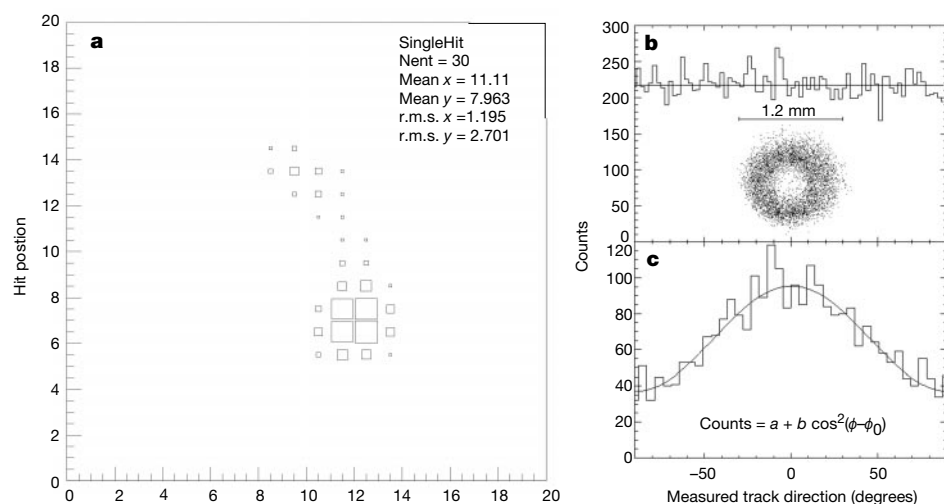


Figure 3 The polarization angle measurement. **a**, Image of a real photoelectron track detected by an MPGC filled with neon (80%) and dimethylether (20%). Scale unit is pixel number and larger boxes correspond to a larger energy loss. Each pixel is 200 μm wide. It is possible to recognize the beginning of the track with the Auger electron followed by the weaker ionization loss of the photoelectron (top) and the end of the track with a much larger energy loss (70% of the total charge for this specific event, bottom). For each photoelectron it is, therefore, possible to measure the initial direction of the track, which carries the memory of the polarization. By reconstructing the impact point of the photon, the real position resolution is much better than that imposed by the track extension and is only a factor of two or three worse than a charge-coupled device (CCD). **b**, **c**, Histograms

of the emission angles of the photoelectron in the detector plane as reconstructed from data like those of **a**, **b**. Unpolarized photons from a Fe⁵⁵ source. No preference in the track direction results in a histogram which is consistent with a flat curve. The loci of the baricentres for a 5.4-keV pencil beam of unpolarized photons are displayed, showing how tracks retain their energy loss asymmetry. **c**, Nearly 100% polarized photons from a 5.4-keV extended source. The amplitude of the cos² fit to the histogram of counts is directly related to the sensitivity of a real polarimeter. The angular phase is the direction of polarization of the incoming photons. The so called modulation factor $(C_{\text{max}} - C_{\text{min}})/(C_{\text{max}} + C_{\text{min}})$ is measured to be 0.44. The χ^2_{red} to the fit is 1.02; $a = 36.98 \pm 1.84$; $b = 58.43 \pm 3.57$; $\phi_0 = 0.45^\circ \pm 1.73^\circ$.

the theoretical distribution of the photoelectron and the smearing due to scattering.

We chose a Ne-based gas mixture to fill the gas chamber because in the energy band of interest the photoelectron track is longer and fires several pixels, while retaining reasonable efficiency. A low-atomic-number gas is less efficient for detection of primary photons. However, the scattering/slowing ratio is lower as well: the track is straighter and the direction of emission can be measured more precisely. Moreover the Ne K-edge energy is so low (0.87 keV) that the accompanying isotropic Auger electron does not blur the information on the polarization; in fact, it makes the identification of the impact point easier. The use of even lower K-edge converters together with a very fine pixel size could make low-energy polarimeters (0.5–2 keV) possible.

With our prototype we have demonstrated the practical feasibility of a new generation of photoelectric polarimeters in the 2–10 keV band. The device can simultaneously also produce good images (50–100 μm), moderately good spectroscopy (16% full-width at half-maximum, at 5.4 keV), and fast, high-rate timing down to 150 eV. Moreover, being truly two-dimensional, it is non-dispersive and does not require rotation.

We also tested our capability to model the polarization detection processes. As absorption, slowing down, scattering and transverse diffusion of electrons in the drift are well known quantities, we may reliably predict the performance of another detector configuration that would be better optimized for astrophysical applications. It is based on an existing³¹ VLSI (Very Large Scale Integration) readout chip combined with other well established detector technology. We can derive the polarimetric sensitivity of such detectors when installed at the focus of a real X-ray telescope. In Table 1 we compare the sensitivity of the present and final configuration of the MPGC with SXP.

The MPGC requires integration periods that are about 100 times shorter than those of the SXP to detect the same polarization in bright sources. With integrations of the order of one day we could perform polarimetry of active galactic nuclei at the 1% level, a breakthrough in this fascinating window of high-energy astrophysics.

sics. With the planned XEUS³² telescope, polarimetry could become a new high-throughput branch of X-ray astronomy. □

Received 23 January; accepted 2 April 2001.

- Mészáros, P., Novick, R., Chanan, G. A., Weisskopf, M. C. & Szentgyörgyi, A. Astrophysical implication and observational prospects of X-ray polarimetry. *Astrophys. J.* **324**, 1056–1067 (1988).
- Rees, M. J. Expected polarization properties of binary X-ray sources. *Mon. Not. R. Astron. Soc.* **171**, 457–465 (1975).
- Sunyaev, R. A. & Titarchuk, L. G. Comptonization of low-frequency radiation in accretion disks: angular distribution and polarization of hard radiation. *Astron. Astrophys.* **143**, 374–388 (1985).
- Gnedin, Yu. N., Pavlov, G. G. & Shibano, Yu. A. The effect of vacuum birefringence in a magnetic field on the polarization and beaming of X-ray pulsars. *Sov. Astron. Lett.* **4**, 117–119 (1978).
- Ventura, J. Scattering of light in strongly magnetized plasma. *Phys. Rev. D* **19**, 1684–1695 (1979).
- Mészáros, P. & Ventura, J. Vacuum polarization effects on radiative opacities in strong magnetic field. *Phys. Rev. D* **19**, 3565–3575 (1979).
- Cash, W., Shapley, A., Osterman, S. & Marshall, J. Laboratory detection of X-ray fringes with grazing-incidence interferometer. *Nature* **407**, 160–162 (2000).
- Gnedin, Yu. N. & Sunyaev, R. A. Polarization of optical and X-ray radiation from compact thermal sources with magnetic field. *Astron. Astrophys.* **36**, 379–394 (1974).
- Connors, P. A. & Stark, R. F. Observable gravitational effects on polarized radiation coming from near a black hole. *Nature* **269**, 128–129 (1977).
- Stark, R. F. & Connors, P. A. Observational test for the existence of a rotating black hole in Cyg X-1. *Nature* **266**, 429–430 (1977).
- Connors, P. A., Piran, T. & Stark, R. Polarization features of X-ray radiation emitted near a black hole. *Astrophys. J.* **235**, 224–244 (1980).
- Bao, G., Wiita, P. & Hadrava, P. Energy-dependent polarization variability as a black hole signature. *Phys. Rev. Lett.* **77**, 12–15 (1996).
- Celotti, A. & Matt, G. Polarization properties of synchrotron self-Compton emission. *Mon. Not. R. Astron. Soc.* **268**, 451–458 (1994).
- Poutanen, J. Relativistic jets in blazars: polarization of radiation. *Astrophys. J. Suppl. Ser.* **92**, 607–609 (1994).
- Tanaka, Y. et al. Gravitationally redshifted emission implying an accretion disk and massive black-hole in the active galaxy MCG-6-30-15. *Nature* **375**, 659–661 (1995).
- Ogura, J., Nobuyori, O. & Kojima, Y. Profiles and polarization properties of emission lines from relativistic disks. *Publ. Astron. Soc. Jpn* **52**, 841–845 (2000).
- Matt, G., Fabian, A. C. & Ross, R. R. X-ray photoionized accretion discs: UV and X-ray continuum spectra and polarization. *Mon. Not. R. Astron. Soc.* **264**, 839–852 (1993).
- Novick, R., Weisskopf, M. C., Berthelsdorf, R., Linke, R. & Wolff, R. S. Detection of X-ray polarization of the Crab Nebula. *Astrophys. J.* **174**, L1–L8 (1972).
- Weisskopf, M. C., Silver, E. H., Kestenbaum, H. L., Long, K. S. & Novick, R. A precision measurement of the X-ray polarization of the Crab Nebula without pulsar contamination. *Astrophys. J.* **220**, L117–L122 (1978).
- Kaaret, P. et al. SXP: a focal plane stellar X-ray polarimeter for the Spectrum-X-Gamma mission. *Opt. Eng.* **29**, 773–783 (1990).
- Schnopper, H. W. & Kalata, K. Polarimeter for celestial X-rays. *Astron. J.* **74**, 854–858 (1969).
- Novick, R. in *Planets, Stars and Nebulae Studied with Photopolarimetry* (ed. Gerehls, T.) 262–317 (Univ. Arizona Press, Tucson, 1972).

23. Riegler, G. R., Garmire, G. P., Moore, W. E. & Stevens, J. A low-energy X-ray polarimeter. *Bull. Am. Phys. Soc.* **15**, 635 (1970).
24. Tsunemi, H. *et al.* Detection of X-ray polarization with a charge coupled device. *Nucl. Instrum. Methods A* **321**, 629–631 (1992).
25. Buschhorn, G. *et al.* X-ray polarimetry using the photoeffect in a CCD detector. *Nucl. Instrum. Methods A* **346**, 578–588 (1994).
26. Soffitta, P. *et al.* Astronomical X-ray polarimetry based on photoelectric effect with microgap detectors. *Nucl. Instrum. Methods A* (in the press); also preprint astro-ph/0012183 at (xxx.lanl.gov) (2000).
27. Angelini, F. *et al.* The micro-gap chamber. *Nucl. Instrum. Methods A* **335**, 69–77 (1993).
28. Austin, R. A., Minamitani, T. & Ramsey, B. Development of a hard X-ray imaging polarimeter. *Proc. SPIE* **2010**, 118–125 (1993).
29. La Monaca, A. *et al.* A new photoelectron imager for X-ray astronomical polarimetry. *Nucl. Instrum. Methods A* **416**, 267–277 (1998).
30. Sauli, F. GEM: a new concept for electron amplification in gas detectors. *Nucl. Instrum. Methods A* **386**, 531–534 (1997).
31. Campbell, M. *et al.* A pixel readout chip for 10–30 Mrad in standard 0.25 μm CMOS. *IEEE Trans. Nucl. Sci.* **46**, 156–160 (1999).
32. Bavdaz, M. *et al.* Status of the X-ray evolving universe spectroscopy mission (XEUS). *Proc. SPIE* **4138**, 69–78 (2000).
33. Christensen, F. E. *et al.* X-ray calibration of the SODART flight telescope. *Proc. SPIE* **3113**, 69–78 (1997).

Acknowledgements

This work is partially supported by the Italian Space Agency (ASI).

Correspondence and requests for materials should be addressed to E.C. (e-mail: costa@ias.rm.cnr.it).

Fabry–Perot interference in a nanotube electron waveguide

Wenjie Liang^{*†}, Marc Bockrath^{†‡}, Dolores Bozovic[‡], Jason H. Hafner^{*}, M. Tinkham[‡] & Hongkun Park^{*}

^{*} Department of Chemistry and Chemical Biology and [‡] Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

[†] These authors contributed equally to this work

The behaviour of traditional electronic devices can be understood in terms of the classical diffusive motion of electrons. As the size of a device becomes comparable to the electron coherence length, however, quantum interference between electron waves becomes increasingly important, leading to dramatic changes in device properties^{1–8}. This classical-to-quantum transition in device behaviour suggests the possibility for nanometer-sized electronic elements that make use of quantum coherence^{1,2,7,8}. Molecular electronic devices are promising candidates for realizing such device elements because the electronic motion in molecules is inherently quantum mechanical^{9,10} and it can be modified by well defined chemistry^{11–13}. Here we describe an example of a coherent molecular electronic device whose behaviour is explicitly dependent on quantum interference between propagating electron waves—a Fabry–Perot electron resonator based on individual single-walled carbon nanotubes with near-perfect ohmic contacts to electrodes. In these devices, the nanotubes act as coherent electron waveguides^{14–16}, with the resonant cavity formed between the two nanotube–electrode interfaces. We use a theoretical model based on the multichannel Landauer–Büttiker formalism^{17–19} to analyse the device characteristics and find that coupling between the two propagating modes of the nanotubes caused by electron scattering at the nanotube–electrode interfaces is important.

Isolated single-walled nanotubes (SWNTs) were synthesized by chemical vapour deposition¹³, and electrical devices based on individual SWNTs were fabricated as reported previously using electron-beam lithography¹³. More than 100 nanotube devices have

been made and classified as metallic or semiconducting on the basis of their resistance versus gate voltage (V_g) behaviour^{13,20}. Most metallic nanotube devices exhibited room-temperature resistance below 100 k Ω , and more than 20 devices exhibited resistance below 15 k Ω . The lowest resistance values observed in some metallic nanotube devices were around 7 k Ω , approaching the theoretical lower limit of 6.5 k Ω for a nanotube device with perfect ohmic contacts^{2,14,21}. This observation indicates that the contacts between nanotubes and the Au/Cr electrodes in our nanotube devices are nearly perfect, unlike those in previous SWNT devices^{13,16,21–24}, and that electrons can pass through the nanotube–metal junction with little reflection. The following discussion concentrates on metallic nanotube devices with room-temperature resistances below 15 k Ω .

Figure 1 shows a differential conductance ($\partial I/\partial V$)– V_g plot near zero bias ($V = 0$) obtained from a representative nanotube device at a temperature $T = 4$ K. The length of the nanotube segment (L) between two electrodes was around 200 nm, as determined by atomic force microscopy. Below $T = 10$ K, the device exhibits pronounced $\partial I/\partial V$ oscillations that are quasi-periodic in V_g , with an average conductance of around $3.2 e^2/h$ (the value of e^2/h is 38.8 μS or $(25.8 \text{ k}\Omega)^{-1}$). Figure 2 shows two-dimensional $\partial I/\partial V$ plots as a function of V and V_g obtained from nanotube devices with $L \approx 530$ nm and $L \approx 220$ nm, respectively. The dips in $\partial I/\partial V$ appear as dark lines in Fig. 2. The positions of the $\partial I/\partial V$ dips evolve smoothly as V and V_g change, forming a mesh of crisscrossing dark lines. Similar $\partial I/\partial V$ – V – V_g patterns were observed in 10 other devices with average conductance above $2e^2/h$, although the V and V_g spacing between adjacent dark lines changed from device to device. The conductance behaviour of these devices did not change substantially as the temperature was reduced from 4 K to 100 mK.

Figures 1 and 2 illustrate several characteristics shared by all 12 nanotube devices that exhibit $\partial I/\partial V$ oscillations. The average values of $\partial I/\partial V$ were around 2 – $3 e^2/h$, and $\partial I/\partial V$ remained above e^2/h irrespective of V , V_g and T , clearly indicating that the electrical behaviour of these nanotube devices is distinct from those reported

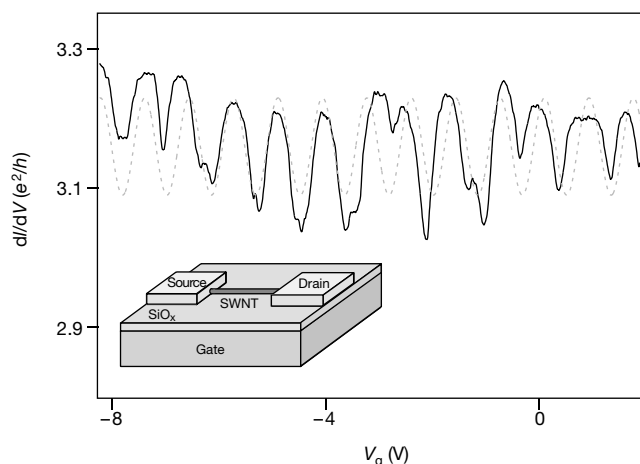


Figure 1 Zero-bias differential conductance ($\partial I/\partial V$) of a 200-nm SWNT device plotted against gate voltage (V_g). Isolated SWNTs were synthesized on a degenerately doped silicon wafer with a 1- μm oxide layer by chemical vapour deposition. Individual SWNTs with ~ 1 -nm height were located by atomic force microscopy, and nanotube devices were fabricated by defining two Au/Cr electrodes on top of the SWNTs by electron-beam lithography. Electrical properties of nanotube devices were characterized as a function of bias voltage (V) and V_g . The degenerately doped silicon substrate was used as a gate electrode to modulate the charge density and the Fermi-level position within the nanotubes. The dotted curve shows a sinusoidal function with the same average period as the measured data. Comparison between these two plots shows that the measured data is quasi-periodic in V_g . Inset, a schematic diagram of the SWNT device, showing a nanotube with attached leads, the insulating gate oxide and the degenerately doped silicon gate.

previously that exhibited a Coulomb-blockade behaviour^{13,16,21–25}. In addition, the $\partial I/\partial V$ dips were typically more pronounced than $\partial I/\partial V$ peaks, as shown by the dominant dark lines in Fig. 2. Most importantly, the V and V_g spacing between adjacent dark lines increased as L decreased. This last behaviour is best quantified by the bias voltage (V_c) at the crossing point between adjacent left- and right-sloped dark lines: inspection of Fig. 2 shows that V_c for the 530-nm device is around 3.5 mV, whereas V_c for the 220-nm device is around 6.5 mV. The inset in Fig. 2 shows the plot of V_c against L^{-1} obtained from seven different devices where V_c can be unambiguously determined, and it clearly shows the linear relationship between V_c and L^{-1} .

Previous experimental^{15,21–28} and theoretical^{14,21} studies have indicated that metallic SWNTs behave as a one-dimensional ballistic conductor where the current is carried by two spin-degenerate one-dimensional transport modes with linear dispersion (Fig. 3). The maximum $\partial I/\partial V$ value near zero bias expected for a SWNT device is therefore $4e^2/h$, obtained only when electrons pass through the nanotube–metal interfaces without reflection^{2,14}. The mean conductance values in Figs 1 and 2 are smaller than this theoretical maximum, indicating electron scattering in the nanotube devices. The linear relationship between V_c and L^{-1} in Fig. 2 (inset) provides experimental evidence that the electron scattering occurs mostly at the nanotube–metal interface and that electrons pass through the nanotube ballistically.

This picture of interfacial electron scattering coupled with ballistic electron transport within the nanotube indicates that the nanotube-device characteristics may be determined by quantum interference between electron waves multiply reflected between two nanotube–metal interfaces, analogous to the light transmission in an optical Fabry–Perot cavity²⁹. Figures 1 and 2 show that $\partial I/\partial V$ oscillates as a function of V_g and V , and the oscillation period exhibits an inverse dependence on the nanotube length. Because V_g modulates the Fermi-level position in the nanotube and hence changes the Fermi wave number ($2\pi/\text{wavelength}$) k for electrons, the observed $\partial I/\partial V$ oscillations can be attributed to the change in k , just as in the case of transmitted

intensity oscillations in an optical cavity.

The observed electron-wave interference can be understood quantitatively on the basis of a theoretical model shown schematically in Fig. 3. In this model, which is based on the multichannel Landauer–Büttiker formalism^{2,17–19}, the nanotube is considered as a coherent electron waveguide with two propagating modes, and the electron scattering is modelled by 4×4 scattering (S) matrices at each interface, S_L and S_R . Following previous theoretical studies^{14,21}, the electron scattering between the two modes inside the nanotube is ignored, and the phase accumulation during electron propagation is represented by a diagonal 4×4 matrix S_N . Inter-mode coupling can still occur through the electron scattering at the interfaces, and the matrix elements of S_L and S_R represent not only the electron transmission and reflection within the same modes but also the coupling between the two propagation modes.

One important difference between the nanotube electron cavity and a simple single-mode optical cavity arises from the fact that the two propagating modes in SWNTs are characterized by different wave vectors, k_1 and k_2 (refs 14, 21). In an isolated neutral SWNT, k_1 and k_2 are the same and they are both given by k_0 ($k_0 = 8.44 \text{ nm}^{-1}$). In our nanotube devices, however, k_1 and k_2 are in general different owing to the linear band dispersion in SWNTs, and the difference between k_1 and k_2 grows linearly as the Fermi level is shifted by V_g . Consequently, electrons in the two propagating modes acquire different phase shifts as they traverse the nanotube, which are represented by the diagonal matrix elements of S_N . The phase change as a function of electron energy is responsible for the interference patterns as a function of V and V_g in Figs 1 and 2.

The calculation of the $\partial I/\partial V$ patterns requires knowledge of the matrix elements of S_L and S_R , each containing 16 independent parameters in general. To reduce the number of parameters in the theoretical fit to the experimental data, most calculations were performed in the reduced parameter space. Specifically, both S_L and S_R were defined by exponentiating a hermitian matrix that contains just four parameters, r_1 , r_2 , δ_1 and δ_2 (see Methods). The resulting S matrices exhibit the required unitarity and properties

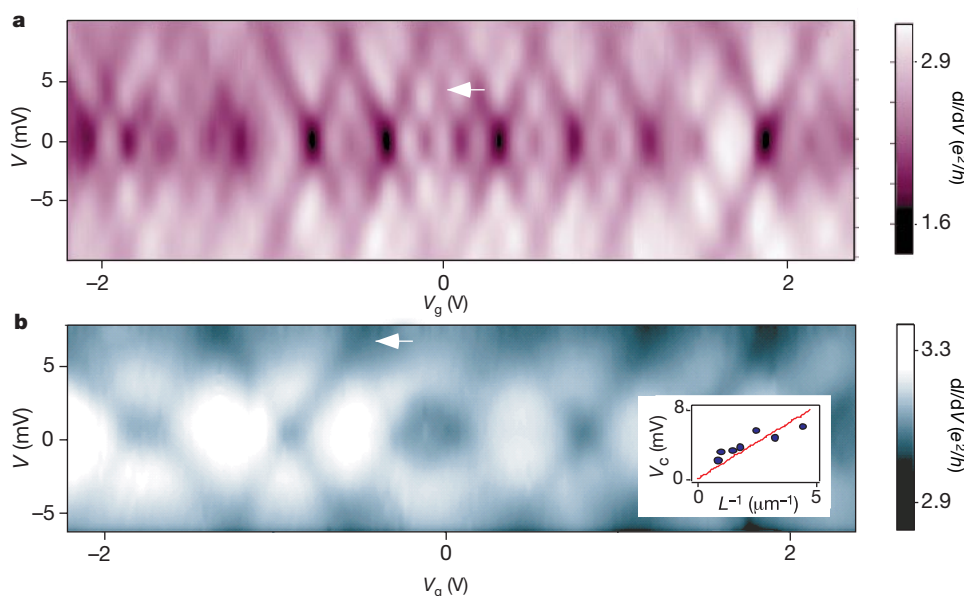


Figure 2 Two-dimensional $\partial I/\partial V$ plots as a function of V and V_g measured at $T = 4 \text{ K}$. **a**, Data from a 530-nm SWNT device; **b**, data from a 220-nm SWNT device. Both plots show a quasi-periodic pattern of crisscrossing dark lines that correspond to the $\partial I/\partial V$ dips as V and V_g are varied. The bias voltage values (V_c) at which adjacent positively and negatively sloped lines intersect (white arrows) quantify the energy scales for $\partial I/\partial V$

oscillations. In **a**, V_c is $\sim 3.5 \text{ meV}$; in **b**, V_c is $\sim 6.5 \text{ meV}$. Inset, values of V_c from seven devices plotted against the inverse nanotube length (L^{-1}). The solid curve is a line with a slope equal to $\hbar v_F/2 = 1,670 \text{ meV nm}^{-1}$, where $v_F = 8.1 \times 10^5 \text{ m s}^{-1}$ is the Fermi velocity in the nanotube.

demanding by time-reversal symmetry. Here r_1 and r_2 correspond to (up to a constant) the magnitude of the electron reflection matrix elements in the first-order Born approximation with and without the inter-mode hopping, while δ_1 and δ_2 signify the phases of these matrix elements¹⁹.

The calculations were further simplified by assuming that S_L and S_R are mirror images of each other. This assumption is motivated by the experimental observation that the $V - V_g$ slopes of right- and left-sloping $\partial I/\partial V$ lines are nearly identical and thus the electron reflection probabilities at the two interfaces are approximately the same in most devices³⁰. S_L and S_R were also assumed to be energy independent over the energy range accessed in the experiment. With

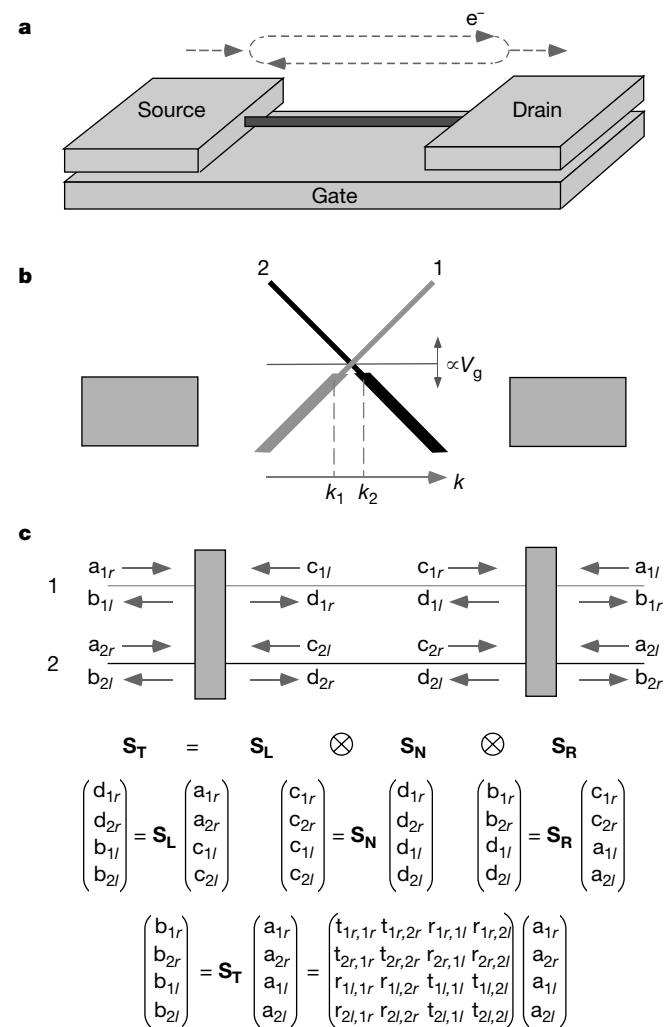


Figure 3 A theoretical model that explains the observed interference patterns. **a**, Schematic diagram of a SWNT device illustrating the multiple electron reflection that gives rise to the observed interference pattern. **b**, Diagram of the band dispersion relation of a metallic SWNT near a band-crossing point. In SWNTs, electronic states near two Fermi points located at $\vec{k} = \vec{k}_0$ and $\vec{k}' = -\vec{k}_0$ contribute to the electrical conduction. The diagram shown here depicts only the band dispersion relation near $\vec{k}$ for clarity. At the $\vec{k}$ ($\vec{k}'$) point, two bands derived from the bonding (π) and antibonding (π^*) orbitals between neighbouring carbon atoms cross with equal slopes of opposite sign. The thick lines represent the electron filling in each band. The diagram shows the effect of V_g on the electron filling in SWNTs when V is zero. The shift of the band edge (or the band-crossing point) with V_g results in a change in the charge density in the nanotube, which, in turn, changes the Fermi wave numbers k_1 and k_2 . **c**, A diagram illustrating the basis wave functions used for defining the S-matrices S_L , S_N , S_R and S_T . The basis-set ordering was chosen so that the total scattering matrix S_T is a unit diagonal matrix in the absence of interfacial scattering.

these approximations, the theoretical fit to the experimental data can be performed using four parameters.

The overall conductance behaviour of a nanotube device can be calculated once the device S matrix S_T is obtained from S_L , S_R and S_N by matrix combination^{2,17–19}. Specifically, in the zero-temperature limit, $\partial I/\partial V$ as a function of V and V_g in a nanotube device is related to the matrix elements of S_T as (see Methods for derivation)

$$\frac{\partial I}{\partial V}(V, V_g) = \frac{2e^2}{h} \left[\sum_{i,j=1,2} \left| t_{il,jr} \left(\frac{eV}{2\hbar v_F} + \frac{\pi C_L V_g}{4e} \right) \right|^2 + \sum_{i,j=1,2} \left| t_{il,jr} \left(\frac{-eV}{2\hbar v_F} + \frac{\pi C_L V_g}{4e} \right) \right|^2 \right]$$

In this formula, $t_{il,jr}$ represents matrix elements of S_T defined in Fig. 3, v_F is the Fermi velocity inside the nanotube ($v_F \approx 8 \times 10^5 \text{ m s}^{-1}$), and C_L is the capacitance per unit length of a nanotube. This formula specifies not only the linear-response behaviour of a nanotube device but also its nonlinear response upon the application of the bias voltage. Although the above formula is derived for the zero-temperature limit, it can still be used to fit the data in Fig. 2 because the observed conductance behaviour did not change as the temperature was lowered from 4 K to 100 mK and hence thermal smearing had little role in determining the transport properties.

The result of the theoretical fit for the 220-nm nanotube device is presented in Fig. 4 along with the experimental data. The theoretical fit clearly accounts for the major features in the data satisfactorily. Although they are not shown, the calculations reproduced experimental data from other devices as well. One of the most important successes of the model is that it explains the energy scale of the observed $\partial I/\partial V$ oscillations (represented by eV_c) without any adjustable parameters. As shown in Fig. 3, two bands in SWNTs have $E-k$ slopes with equal magnitude and opposite sign. This band structure guarantees that whenever the Fermi-level change causes the additional round-trip phase shift of $+2\pi$ in one mode, the other mode acquires a phase shift of -2π . The value of V_c can therefore be obtained by setting the round trip phase shift $2LeV_c/(\hbar v_F)$ equal to 2π . Figure 2 (inset) shows a red line obtained from this expression along with experimental data. It shows that the theoretical model explains experimental V_c values very well without any adjustable parameters.

Figure 4 also shows that the theoretical model can explain the predominance of $\partial I/\partial V$ dips over the peaks in the data. Inspection of the fitting parameters, r_1 , r_2 , δ_1 and δ_2 , shows that the pronounced $\partial I/\partial V$ dips cannot be reproduced without the inter-mode coupling at the interface (represented by r_1 in our fit), especially when the average conductance of the device is above $3e^2/h$ as in Fig. 4. This observation strongly suggests that the inter-mode coupling at the nanotube-metal interface is important in our device geometry. To test whether this conclusion is dependent on the number of fitting parameters, a more extensive model fit was performed, with S_L and S_R containing more than four parameters. The results confirm that the inter-mode coupling is indeed required to explain the dominance of $\partial I/\partial V$ dips.

Despite the good agreement between theory and experiment shown in Fig. 4, several features in the experimental data remain unexplained by the theoretical model. Most notably, Fig. 1 shows that the magnitudes of the observed $\partial I/\partial V$ dips are not uniform and show variations on the order of $0.1e^2/h$ as V_g is varied. Moreover, Fig. 2a shows that this variation leads to superstructures in the two-dimensional plot that are not periodic in V_g . These observations may indicate the possible effects of disorder on electron propagation in the nanotube. This disorder is likely to arise from the underlying substrate or at the interface because intrinsic nanotube defects change electron transport properties much more markedly¹³. Figure 2 also shows that the magnitudes of $\partial I/\partial V$ modulation

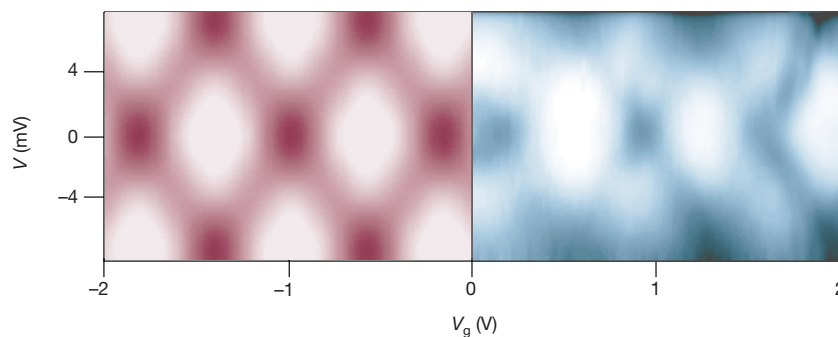


Figure 4 The calculated (left, shown in red) and measured (right, shown in blue) two-dimensional dI/dV plots as a function of V and V_g for a 220-nm SWNT device. These plots have been shifted in V_g for easy comparison. In both plots, dark corresponds to $2.9 e^2/h$,

and white corresponds to $3.2 e^2/h$. The theoretical calculation was performed by setting $r_1 = 0.2$, $r_2 = 0.25$, $\delta_1 = 0.4$, $\delta_2 = 0.95$ for both S_L and S_R and with $C_L \approx 20$ electrons $V^{-1} \mu m^{-1}$ (V in V_g).

diminish as $|V|$ is increased. This observation suggests the occurrence of heating or dephasing when the electron energy distribution deviates from equilibrium. The dephasing may arise from the fact that the nanotube–metal interface is not abrupt and instead spans many atoms. □

Methods

The explicit forms of S_L , S_R and S_N

The S matrices, S_L , S_R and S_N , and the basis wave functions are defined in Fig. 3. Assuming that the electron scattering between the two modes inside the nanotube is ignored^{14,21}, the matrix S_N is given by

$$S_N = \begin{bmatrix} e^{i\phi_1} & 0 & 0 & 0 \\ 0 & e^{i\phi_2} & 0 & 0 \\ 0 & 0 & e^{i\phi_1} & 0 \\ 0 & 0 & 0 & e^{i\phi_2} \end{bmatrix} \quad (1)$$

Here ϕ_1 and ϕ_2 are the phases accumulated by electrons in the two transport modes upon traversing the nanotube segment between the metal electrodes. In general, S_L and S_R each contain 16 independent parameters. To reduce the number of parameters in the theoretical fit of experimental data, S_L and S_R were parameterized on the basis of the first-order Born approximation. Assuming a weak scattering potential Δ at the interface, the scattering matrices at the interface can be obtained using the Born approximation¹⁹

$$S_{mn} \approx \delta_{mn} + ia_{mn} = \delta_{mn} + \frac{i}{\hbar v_F} \langle \psi_m | \Delta | \psi_n \rangle \quad (2)$$

Here δ_{mn} is the Kronecker delta, and ψ_m represents mode wave functions defined in Fig. 3. Note that the S matrix in equation (2) is only unitary to first order in Δ . To make it unitary, the S matrix was taken to be the matrix exponential of $\{a_{mn}\}$: $S = \exp(ia)$. Specifically, the calculation shown in Fig. 4 was performed using the following ansatz for S_R and S_L :

$$S_{R(L)} = \exp \left(i \begin{bmatrix} 0 & 0 & r_2 & r_1 \exp(\pm i\delta_1) \\ 0 & 0 & r_1 \exp(\pm i\delta_1) & r_2 \exp(\pm i\delta_2) \\ r_2 & r_1 \exp(\mp i\delta_1) & 0 & 0 \\ r_1 \exp(\mp i\delta_1) & r_2 \exp(\mp i\delta_2) & 0 & 0 \end{bmatrix} \right) \quad (3)$$

Here r_1 and r_2 are proportional to the magnitude of the electron reflection matrix elements in the first-order Born approximation with and without the inter-mode hopping, respectively, and δ_1 and δ_2 signify the phases of these matrix elements. As described in the text, the definitions of S_R and S_L in equation (3) ensure the opposite parity between the left and right electrodes relative to the nanotube. The scattering matrices S_R and S_L are further assumed to be independent of energy. Therefore, the conductance variation as a function of V and V_g depends only on the phase shifts $\phi_{1,2}$ that appear in S_N .

Derivation of the expression for dI/dV

A non-zero source-drain bias V raises the electrochemical potential of the left electrode by $eV/2$ and lowers the electrochemical potential of the right electrode by $eV/2$. Because the electron transport through the nanotube is ballistic, the total energy of the electrons involved in transport is conserved. The total energy $E(-eV/2 < E < eV/2)$ of electrons at the Fermi surface can be expressed as a sum of their kinetic energy $K(x)$ and the local self-consistent potential energy $\Phi_{ev}(x)$, where x is the spatial position along the nanotube:

$$K(x) + \Phi_{ev}(x) = E \quad (4)$$

The wave numbers $k_{1,2}$ of the electrons in the two transport modes are given by

$$k_{1,2}(x) = k_0 \pm \frac{K(x)}{\hbar v_F} = k_0 \pm \frac{E - \Phi_{ev}(x)}{\hbar v_F} \quad (5)$$

and the resulting phase shifts are given by

$$\phi_{1,2} = \int_0^L k_{1,2}(x) dx = \int_0^L \left(k_0 \pm \frac{E - \Phi_{ev}(x)}{\hbar v_F} \right) dx = k_0 L \pm \frac{E - \langle \Phi_{ev} \rangle}{\hbar v_F} L \quad (6)$$

Here $\langle \Phi_{ev} \rangle$ is the value of the self-consistent potential averaged over the nanotube length L . $\langle \Phi_{ev} \rangle$ can be related to the excess charge Q on the nanotube using the local density of states per unit length $\rho(0) = 8/(\hbar v_F)$ at the Fermi level:

$$Q = - \int_0^L e \Phi_{ev}(x) \rho(0) dx = - \frac{8Le}{\hbar v_F} \langle \Phi_{ev} \rangle \quad (7)$$

From equations (6) and (7), it follows that

$$\phi_{1,2} = k_0 L \pm \frac{E}{\hbar v_F} L \pm \frac{\pi Q}{4e} \quad (8)$$

As S_L and S_R are mirror images of each other, the total charge on the nanotube Q is unaffected by V . Therefore, Q can be related to V_g by $Q = C_L V_g$, where C_L is the capacitance per unit length. The result for $\phi_{1,2}$ can thus be obtained as

$$\phi_{1,2} = k_0 L \pm \left(\frac{E}{\hbar v_F} L + \frac{\pi C_L V_g}{4e} \right) \quad (9)$$

The phases, $\phi_{1,2}$, given by equation (9) determine S_N . The matrix S_T can be obtained by combining S_L , S_R and S_N by matrix combination^{2,17–19}. Once S_T is known, the current through the nanotube can be calculated by

$$I = \frac{2e}{h} \int_{-eV/2}^{eV/2} dE \sum_{i,j=1,2} \left| t_{ij} \left(\frac{E}{\hbar v_F} + \frac{\pi C_L V_g}{4e} \right) \right|^2 \quad (10)$$

where the matrix elements t_{ij} are defined in Fig. 3. The differentiation of equation (10) leads to the expression for dI/dV given in the text.

Received 29 December 2000; accepted 14 March 2001.

- Ando, T. *et al.* *Mesoscopic Physics and Electronics* (Springer, Berlin, 1998).
- Datta, S. *Electronic Transport in Mesoscopic Systems* (Cambridge Univ. Press, Cambridge, 1995).
- van Wees, B. J. *et al.* Observation of zero-dimensional states in a one-dimensional electron interferometer. *Phys. Rev. Lett.* **62**, 2523–2526 (1989).
- Crommie, M., Lutz, C. P. & Eigler, D. M. Confinement of electrons to quantum corrals on a metal surface. *Science* **262**, 218–220 (1993).
- Ji, Y. *et al.* Phase evolution in a Kondo-correlated system. *Science* **290**, 779–783 (2000).
- Topinka, M. A. *et al.* Imaging coherent electron flow from a quantum point contact. *Science* **289**, 2323–2326 (2000).
- Manoharan, H. C., Lutz, C. P. & Eigler, D. M. Quantum mirages formed by coherent projection of electronic structure. *Nature* **403**, 512–515 (2000).
- Debray, P. *et al.* Ballistic electron transport in stubbed quantum waveguides: Experiment and theory. *Phys. Rev. B* **61**, 10950–10958 (2000).
- Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
- Park, H. *et al.* Nano-mechanical oscillations in a single-C₆₀ transistor. *Nature* **407**, 57–60 (2000).
- Chen, J., Reed, M. A., Rawlett, A. M. & Tour, J. M. Large on-off ratios and negative differential resistance in a molecular electronic device. *Nature* **286**, 1550–1552 (1999).
- Collier, C. P. *et al.* A [2]catenane-based solid state electronically reconfigurable switch. *Science* **289**, 1172–1175 (2000).

13. Bockrath, M. *et al.* Resonant electron scattering by defects in single-walled carbon nanotubes. *Science* **291**, 283–285 (2001).
14. White, C. T. & Todorov, T. N. Carbon nanotubes as long ballistic conductors. *Nature* **393**, 240–242 (1998).
15. Frank, S., Poncharal, P., Wang, Z. L. & De Heer, W. A. Carbon nanotube quantum resistors. *Science* **280**, 1744–1746 (1998).
16. Dekker, C. Carbon nanotubes as molecular quantum wires. *Phys. Today* **52** (5), 22–28 (1999).
17. Büttiker, M., Imry, Y., Landauer, R. & Pinhas, S. Generalized many-channel conductance formula with application to small rings. *Phys. Rev. B* **31**, 6207–6215 (1985).
18. Büttiker, M. Role of quantum coherence in series resistors. *Phys. Rev. B* **33**, 3020–3026 (1986).
19. Cahay, M., McLennan, M. & Datta, S. Conductance of an array of elastic scatterers: A scattering-matrix approach. *Phys. Rev. B* **37**, 10125–10136 (1988).
20. Tans, S. J., Verschueren, A. R. M. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393**, 49–52 (1998).
21. McEuen, P. L. *et al.* Disorder, pseudospins, and backscattering in carbon nanotubes. *Phys. Rev. Lett.* **83**, 5098–5101 (1999).
22. Bockrath, M. *et al.* Single-electron transport in ropes of carbon nanotubes. *Science* **275**, 1922–1925 (1997).
23. Tans, S. J. *et al.* Individual single-wall carbon nanotubes as quantum wires. *Nature* **386**, 474–476 (1997).
24. Tans, S. J., Devoret, M. H., Groeneveld, R. J. A. & Dekker, C. Electron–electron correlations in carbon nanotubes. *Nature* **394**, 761–764 (1998).
25. Nygard, J., Cobden, D. H. & Lindelof, P. E. Kondo physics in carbon nanotubes. *Nature* **408**, 342–346 (2000).
26. Venema, L. C. *et al.* Imaging electron wave functions of quantized energy levels in carbon nanotubes. *Science* **283**, 52–55 (1999).
27. Bachtold, A. *et al.* Aharonov–Bohm oscillations in carbon nanotubes. *Nature* **397**, 673–675 (1999).
28. Tsukagoshi, K., Alphenaar, B. W. & Ago, H. Coherent transport of electron spin in a ferromagnetically contacted carbon nanotube. *Nature* **401**, 572–574 (1999).
29. Hecht, E. *Optics* (Addison-Wesley, Reading, 1987).
30. Sohn, L. L., Kouwenhoven, L. P. & Schön, G. *Mesoscopic Electron Transport* (Kluwer, Dordrecht, 1997).

Acknowledgements

We thank C. M. Lieber for providing facilities to synthesize SWNTs; and C. M. Lieber, B. I. Halperin, E. J. Heller and C. Marcus for discussions and advice. This work is supported by NSF and ONR (M.T.) and the Dreyfus Foundation, NSF, and Harvard University (H.P.). J.H.H. was supported by NIH.

Correspondence and requests for materials should be addressed to H.P. (e-mail: HPark@chemistry.harvard.edu).

Non-Fermi-liquid behaviour in $\text{La}_4\text{Ru}_6\text{O}_{19}$

P. Khalifah*, K. D. Nelson†, R. Jin†, Z. Q. Mao†, Y. Liu†, Q. Huang‡, X. P. A. Gao§, A. P. Ramirez§ & R. J. Cava*

* Department of Chemistry and Princeton Materials Institute, Princeton University, Princeton, New Jersey 08540, USA

† Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

‡ NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899; and Department of Materials and Nuclear Engineering, University of Maryland, College Park, Maryland 20742, USA

§ Lucent Technologies, Murray Hill, New Jersey 07974, USA

Understanding the complexities of electronic and magnetic ground states in solids is one of the main goals of solid-state physics. Transition-metal oxides have proved to be particularly fruitful in this regard, especially for those materials with the perovskite structure, where the special characteristics of transition-metal–oxygen orbital hybridization determine their properties. Ruthenates have recently emerged as an important family of perovskites because of the unexpected evolution from high-temperature ferromagnetism in SrRuO_3 to low-temperature superconductivity in Sr_2RuO_4 (refs 1, 2). Here we show that a

ruthenate in a different structural family, $\text{La}_4\text{Ru}_6\text{O}_{19}$, displays a number of highly unusual properties, most notably non-Fermi-liquid behaviour. The properties of $\text{La}_4\text{Ru}_6\text{O}_{19}$ have no analogy among the thousands of previously characterized transition-metal oxides. Instead, they resemble those of $\text{CeCu}_{6-x}\text{Au}_x$ —a widely studied *f*-electron-based heavy fermion intermetallic compound that is often considered as providing the best example of non-Fermi-liquid behaviour. In the ruthenate, non-Fermi-liquid behaviour appears to arise from just the right balance between the interactions of localized electronic states derived from Ru–Ru bonding and delocalized states derived from Ru–O hybridization.

The unusual properties of $\text{La}_4\text{Ru}_6\text{O}_{19}$ are accentuated by comparison to the highly related compound $\text{La}_3\text{Ru}_3\text{O}_{11}$. Both have cubic crystal structures^{3,4} that are a derivative of the KSbO_3 structure type. The structures (Fig 1a and b) contain pairs of edge-shared RuO_6 octahedra, effectively forming Ru_2O_{10} dimers, which are connected to other dimers through corner-shared oxygens. The result is a fully three-dimensional ruthenium–oxygen network that has the same geometry for both compounds. The formal ruthenium oxidation state is +4.33 in both compounds, leading to the same number of available electrons to fill the electronic states in the Ru–O network. Direct metal–metal bonding occurs in $\text{La}_4\text{Ru}_6\text{O}_{19}$, where the Ru–Ru distance within the dimers, 2.49 Å, is unusually short³. No Ru–Ru bonding occurs in $\text{La}_3\text{Ru}_3\text{O}_{11}$ (ref. 4). The presence of distinct metal–metal bonded dimers in $\text{La}_4\text{Ru}_6\text{O}_{19}$ should lead to the existence of localized electronic states in addition to the many delocalized states usually seen in ruthenates due to the strong Ru–O orbital hybridization^{5,6}. This leads to a profound difference in the electronic properties of the two compounds.

Single crystals of $\text{La}_4\text{Ru}_6\text{O}_{19}$ and $\text{La}_3\text{Ru}_3\text{O}_{11}$ were grown by placing 0.2 g of dried La_2O_3 and RuO_2 mixed in the stoichiometric ratios, 1 g of KCl and 0.131 g of KClO_3 in a 2-ml Al_2O_3 crucible. The crucible was sealed in an evacuated quartz tube, and heated at 900 °C for three to four days. The single crystals (up to 0.3 mm in dimension) were removed from the flux and cleaned by washing with distilled water. The large (15 g) polycrystalline powder sample of $\text{La}_4\text{Ru}_6\text{O}_{19}$ that we used in the neutron-diffraction experiments was synthesized in an analogous manner, but without KCl.

The temperature-dependent resistivity for a representative single crystal of $\text{La}_4\text{Ru}_6\text{O}_{19}$ is presented in Fig. 2, as is the normalized resistance of a representative $\text{La}_3\text{Ru}_3\text{O}_{11}$ crystal. A room-temperature resistivity of 2.5 mΩ cm was observed for $\text{La}_4\text{Ru}_6\text{O}_{19}$, slightly lower than the value previously reported for polycrystalline material⁷. This resistivity is significantly larger than the

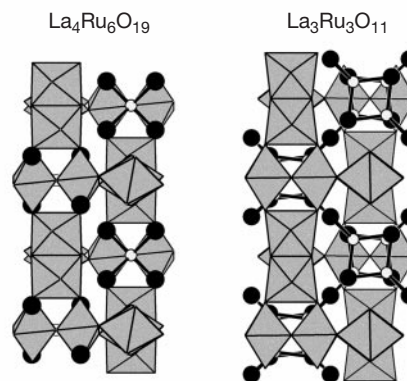


Figure 1 Two crystallographic unit cells of $\text{La}_4\text{Ru}_6\text{O}_{19}$ and $\text{La}_3\text{Ru}_3\text{O}_{11}$. RuO_6 octahedra are shown as shaded polyhedra (with Ru atoms inside). The virtually identical Ru–O network is seen for each. La and non-network O atoms are shown by filled and open circles, respectively.

* Present address: Condensed Matter and Thermal Physics Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

$\sim 0.03 \text{ m}\Omega \text{ cm}$ typically observed for RuO_2 at 300 K (ref. 8). The resistivity ratio $R_{300\text{K}}/R_{1\text{K}}$ is 192 for the $\text{La}_4\text{Ru}_6\text{O}_{19}$ crystal, and $R_{300\text{K}}/R_{4\text{K}}$ is 33 for the $\text{La}_3\text{Ru}_3\text{O}_{11}$ crystal. The shape of the $R(T)$ curve for $\text{La}_3\text{Ru}_3\text{O}_{11}$ is typical for metallic transition-metal oxides. The resistivity of $\text{La}_4\text{Ru}_6\text{O}_{19}$ at low temperatures, however, is unusual—a linear dependence on temperature is observed between approximately 2 and 25 K. In addition, the linear resistivity extrapolates to zero at zero temperature (inset to Fig. 2). This type of resistivity behaviour is observed in the copper oxide superconductors over a wide temperature range, and is taken as one of the primary indications of non-Fermi-liquid behaviour in those materials and in heavy fermion compounds such as $\text{CeCu}_{6-x}\text{Au}_x$ (refs 9–11).

Figure 3 shows the measured magnetic susceptibilities, $\chi(T)$, for collections of small crystals of $\text{La}_4\text{Ru}_6\text{O}_{19}$ and $\text{La}_3\text{Ru}_3\text{O}_{11}$. There is no difference between field-cooled and zero-field-cooled measurements. The data for $\text{La}_3\text{Ru}_3\text{O}_{11}$ (shown in the upper right inset) are typical of those observed in metallic conducting oxides: Pauli paramagnetism from the conduction electrons dominates, with a weakly increasing susceptibility at low temperatures, probably due to a small amount (1–2%) of $S = 1$ paramagnetic impurities. The $\chi(T)$ data (upper data curve in Fig. 3) for $\text{La}_4\text{Ru}_6\text{O}_{19}$ also suggest two contributions: a broad high-temperature peak near 200 K, and a ‘Curie tail’-type contribution at low temperatures. A minimum in $\chi(T)$ is clearly seen at 50 K. The behaviour of $\chi(T)$ below 15 K can be described within the Curie–Weiss framework with an effective Weiss temperature (Θ_W) of -14 K and an effective moment of $0.41 \mu_B$ per Ru atom. The low-temperature susceptibility can alternatively be interpreted within the context of the behaviour observed in non-Fermi-liquid systems such as CeCu_{6-x} (refs 9–11). The upper left inset to Fig. 3 shows the low-temperature $\chi(T)$ data for $\text{La}_4\text{Ru}_6\text{O}_{19}$, with the solid line showing a fit to the $\alpha(1 - \beta T^{1/2})$ dependence often observed in such systems ($\alpha = 1.59 \times 10^{-3}$, $\beta = 0.146$). This interpretation of the low-temperature susceptibility is consistent with the observed non-Fermi-liquid behaviour seen in the transport and specific heat data. Detailed analysis of the low-temperature $\chi(T)$ data is restricted, however, as

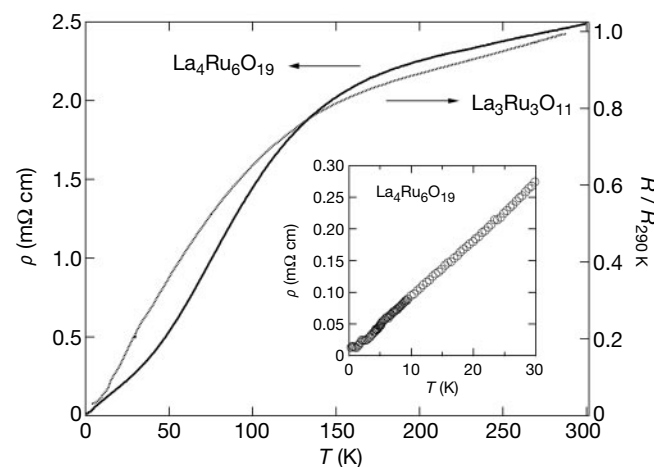


Figure 2 Temperature-dependent resistivity measurements. Temperature-dependent resistivity for a single crystal of $\text{La}_4\text{Ru}_6\text{O}_{19}$ (circles) and the relative resistance for a crystal of $\text{La}_3\text{Ru}_3\text{O}_{11}$ (dotted line) are shown. Inset, the low-temperature resistivity behaviour for $\text{La}_4\text{Ru}_6\text{O}_{19}$. Temperature-dependent d.c. transport measurements were carried out using a Keithley current source and nanovoltmeter in a ^3He refrigerator. True four-probe resistivity measurements were obtained on $\text{La}_4\text{Ru}_6\text{O}_{19}$ by attaching the current and voltage leads with silver epoxy. The small size of the $\text{La}_3\text{Ru}_3\text{O}_{11}$ crystals required that both the current and voltage leads were attached via the same drop of Ag epoxy. These are not true four-point measurements and the data for $\text{La}_3\text{Ru}_3\text{O}_{11}$ are therefore discussed only qualitatively.

there may be some impurity contribution present.

The second contribution to $\chi(T)$ for $\text{La}_4\text{Ru}_6\text{O}_{19}$ can be seen by subtracting the Curie–Weiss fit to the low-temperature data from the observed susceptibility. This contribution decreases with decreasing temperature (lower data curve in Fig. 3). It resembles the behaviour of materials with a gap in the spin excitation spectrum between a non-magnetic ground state and magnetic excited states, such as materials with spin chains and the layered copper oxides¹². If this is the case for $\text{La}_4\text{Ru}_6\text{O}_{19}$, then a generalized

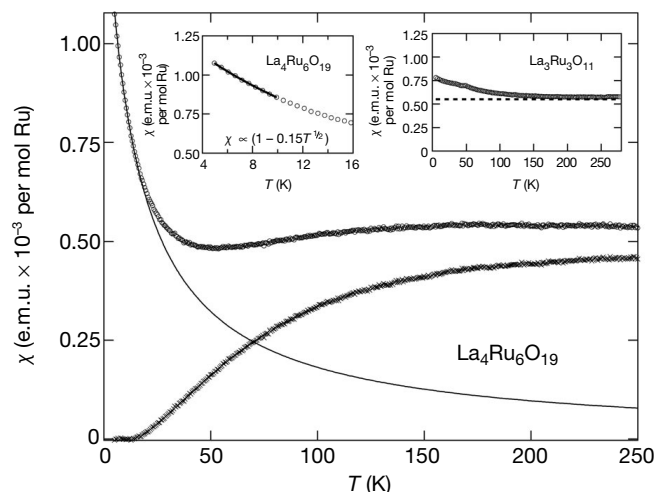


Figure 3 Magnetic susceptibility measurements. The upper data curve shows the measured magnetic susceptibility of $\text{La}_4\text{Ru}_6\text{O}_{19}$. The solid line shows the temperature dependence of the sharply increasing low-temperature $\chi(T)$ fit to the Curie–Weiss law. The lower data curve shows the temperature dependence of the second component of $\chi(T)$, which decreases from approximately 100 K. The solid line through these points at low temperatures shows the potential fit to the generalized 3D spin gap equation. The upper right inset shows $\chi(T)$ (circles) for $\text{La}_3\text{Ru}_3\text{O}_{11}$. The dashed line indicates the temperature-independent paramagnetic contribution. The upper left inset shows the potential fit to the low-temperature $\chi(T)$ data (circles) for $\text{La}_4\text{Ru}_6\text{O}_{19}$ to the $\alpha(1 - \beta T^{1/2})$ behaviour (solid line) often observed for non-Fermi liquids. Magnetic susceptibility measurements were performed on collections of single crystals on the Magnetic Property Measurement System (Quantum Design) in a field of 1 T.

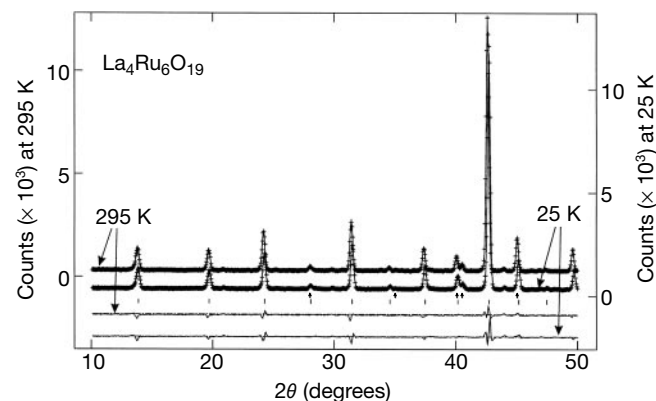


Figure 4 Low-angle powder neutron diffraction data for $\text{La}_4\text{Ru}_6\text{O}_{19}$ at 298 K and 25 K indicating the absence of three-dimensional magnetic ordering at low temperatures. Tick marks indicate calculated $\text{La}_4\text{Ru}_6\text{O}_{19}$ peak positions, while small arrowheads mark the peaks due to a small amount of RuO_2 impurity. The differences between the observed and calculated intensities are shown in the lower traces for each temperature. Experiments were performed at the NIST Center for Neutron Research using the BT-1 32-counter high-resolution powder diffractometer.

three-dimensional gap equation¹³, $\chi(T) \propto T^{1/2} \exp(-\Delta/T)$, can be used to fit the data below 50 K, yielding a gap of $\Delta = 55$ K.

An alternative explanation for a temperature-decreasing $\chi(T)$ component in $\text{La}_4\text{Ru}_6\text{O}_{19}$ might be magnetic ordering. To test for this possibility, neutron-diffraction data were collected to probe the magnetic state of $\text{La}_4\text{Ru}_6\text{O}_{19}$ at low temperatures. A comparison of the low-temperature (25 K) and room temperature (298 K) diffraction data is shown in Fig. 4. No new peaks are present, such as would result from three-dimensional antiferromagnetic ordering. Furthermore, the relative peak intensities are virtually identical at both temperatures, indicating negligible magnetic contributions to the Bragg peaks at low temperatures. The 25 K data can be fitted well to the 298 K structural model³ with no magnetic scattering and no significant changes in the Ru–Ru distances or Ru–O distances and bond angles. Statistical analysis indicates that any ordered moment present would have to be very small—less than $0.04\mu_B$ —to be unobservable in these experiments. Although thermal excitation across a gap in a spin excitation spectrum can be used to describe the susceptibility data, further work will clearly be needed to conclusively determine the microscopic origin of the broadly decreasing component of $\chi(T)$ in $\text{La}_4\text{Ru}_6\text{O}_{19}$.

The temperature-dependent specific heat, plotted as $C(T)/T$ versus T on a logarithmic scale, for various values of applied magnetic field (H), is shown for $\text{La}_4\text{Ru}_6\text{O}_{19}$ in Fig. 5. The inset shows the raw data. For finite values of H , an upturn in $C(T)/T$ is observed at the lowest temperatures, which is probably due to a nuclear Schottky anomaly arising from both La and Ru nuclei. The main panel shows the data with the effects of the nuclear contributions subtracted, by fitting the nuclear Schottky terms to $C(T)/T$ versus T^3 , scaled by H^2 . For Fermi-liquid behaviour at $H = 0$, $C(T)/T$ is expected to be constant at low temperatures. For $\text{La}_4\text{Ru}_6\text{O}_{19}$, however, $C(T)/T$ at $H = 0$ displays $\log(T)$ behaviour below 1 K, approaching a value of about 63 mJ per mol Ru K² as T approaches 0.1 K. The electronic part of the specific heat is suppressed below 2 K by the application of the magnetic field. For applied fields of 1–9 T, the electronic specific heat shows broad peaks near 1 K. At these temperatures, applied magnetic fields suppress spin fluctuations. The H -dependence of $C(T)/T$ therefore indicates the presence of spin fluctuations at low temperatures in $\text{La}_4\text{Ru}_6\text{O}_{19}$. In spite of the underlying fundamental differences

between the chemical and structural characteristics of $\text{La}_4\text{Ru}_6\text{O}_{19}$ and $\text{CeCu}_{6-x}\text{Au}_x$, the behaviour of $C(T)/T$ at low temperatures as a function of applied magnetic fields for these materials is virtually identical, except that the magnitude of $C(T)/T$ is much larger for the intermetallic phase. For $\text{La}_3\text{Ru}_3\text{O}_{11}$ (data not shown) $C(T)/T$ is constant at low temperatures, and independent of applied field (as expected for a Fermi liquid), with $\gamma = 33$ mJ per mol Ru K².

The elementary physical properties of $\text{La}_4\text{Ru}_6\text{O}_{19}$ may be directly compared to those of heavy fermion intermetallic compounds, where there has been extensive consideration of the differences in phenomenology between Fermi-liquid and non-Fermi-liquid behaviour. Unlike the intermetallic compounds, there is no chemical disorder in $\text{La}_4\text{Ru}_6\text{O}_{19}$, which may simplify theoretical considerations. Features such as a low-temperature resistivity varying as T rather than T^2 , an $\alpha(1 - \beta T^{1/2})$ dependence of the low-temperature $\chi(T)$ rather than temperature-independent $\chi(T)$, and a $\log(T)$ dependence of $C(T)/T$ and its magnetic-field dependence rather than temperature- and field-independence have all been observed in intermetallic heavy fermion materials such as $\text{CeCu}_{6-x}\text{Au}_x$ and $\text{U}_{1-x}\text{Y}_x\text{Pd}_3$ (refs 9–11, 14). These phenomena have distinguished them as non-Fermi liquids, which is clearly the case for $\text{La}_4\text{Ru}_6\text{O}_{19}$ as well.

For the f -electron-based intermetallics, the non-Fermi-liquid behaviour at low temperatures has been attributed to electronic correlations owing to the proximity of a zero-temperature (quantum) phase transition occurring at the crossover from a finite temperature, long-range antiferromagnetic ordered ground state to a non-magnetic ground state with short-range magnetic correlations. The competition between those ground states is derived from competition between f -electron local moments interacting antiferromagnetically both directly and through the conduction electrons. This obviously cannot be the case for $\text{La}_4\text{Ru}_6\text{O}_{19}$, and it will be of interest to establish the interactions that lead to its unusual properties. □

Received 11 January; accepted 9 April 2001.

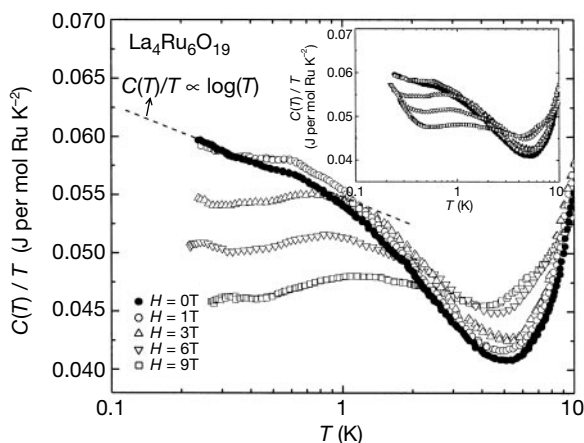


Figure 5 The specific heat data for $\text{La}_4\text{Ru}_6\text{O}_{19}$ corrected for nuclear Schottky anomalies (the inset shows raw data). The $\log(T)$ dependence of the $H = 0$ data at low temperatures is emphasized with a dashed line. The specific heat, $C(T)$, data for both compounds were taken by a standard semiadiabatic heat pulse technique. For both materials, the sample powder was cold-sintered into a hard disk with Ag powder at a mass ratio of about 1:1 to enhance thermalization. The Ag contribution to $C(T)$ was measured separately and subtracted.

1. Longo, J. M., Raccach, P. M. & Goodenough, J. B. Magnetic properties of SrRuO_3 and CaRu_3 . *J. Appl. Phys.* **39**, 1327–1328 (1968).
2. Maeno, Y. *et al.* Superconductivity in a layered perovskite without copper. *Nature* **372**, 532–534 (1994).
3. Abraham, F., Trehoux, J. & Thomas, D. La liaison métal–métal dans les clusters $\text{M}_{12}\text{O}_{36}$: I—préparation et étude structurale des phases $\text{La}_4\text{M}_6\text{O}_{19}$ ($\text{M} = \text{Ru}, \text{Os}$). *Mater. Res. Bull.* **12**, 43–52 (1977).
4. Cotton, F. A. & Rice, C. E. The crystal structure of $\text{La}_3\text{Ru}_3\text{O}_{11}$: A new cubic KSbO_3 derivative oxide with no metal–metal bonding. *J. Solid State Chem.* **25**, 137–142 (1978).
5. Mazin, I. I. & Singh, D. J. Electronic structure and magnetism in Ru-based perovskites. *Phys. Rev. B* **56**, 2556–2571 (1997).
6. Santi, G. & Jarlborg, T. Calculation of the electronic structure and the magnetic properties of SrRuO_3 and CaRuO_3 . *J. Phys. Condens. Matter* **9**, 9563–9584 (1997).
7. Shaplygin, L. S. & Lazarev, V. B. Compounds with the general formula $\text{La}_4\text{M}_6\text{O}_{19}$ —binary oxides of platinum group metals in different oxidation states. *Russ. J. Inorg. Chem.* **241**, 644–646 (1978).
8. Shafer, M. W., Figat, R. A., Olson, B., LaPlaca, S. J. & Argillelo, J. Preparation and characterization of ruthenium dioxide crystals. *J. Electrochem. Soc.* **126**, 1625–1628 (1979).
9. Maple, M. B. *et al.* Non-Fermi liquid ground states in strongly correlated f -electron materials. *J. Low-Temp. Phys.* **99**, 223–249 (1995).
10. von Lohneysen, H. Non-Fermi-liquid behavior in heavy fermion systems. *Physica B* **206** & **207**, 101–107 (1995).
11. von Lohneysen, H. Non-Fermi liquid behavior in the heavy fermion system $\text{CeCu}_{6-x}\text{Au}_x$. *J. Phys. Condens. Matter* **8**, 9689–9706 (1996).
12. Azuma, M., Hiroi, Z., Takano, M., Ishida, K. & Kitaoka, Y. Observation of a spin gap in SrCu_2O_3 comprising spin-1/2 quasi-1D two-leg ladders. *Phys. Rev. Lett.* **73**, 3463–3466 (1994).
13. Sachdev, S. *Quantum Phase Transitions* 95 (Cambridge Univ. Press, Cambridge, UK, 1999).
14. Andraka, B. & Tsvetlik, A. M. Observation of non-Fermi-liquid behavior in $\text{U}_{0.2}\text{Y}_{0.8}\text{Pd}_3$. *Phys. Rev. Lett.* **67**, 2886–2889 (1991).

Acknowledgements

We thank Y. Zadorozhny and L.-N. Zou for help with the low-temperature transport measurements, and Z. Soos and J. Willson for helpful discussions on metal–metal bonding. The work at Princeton University and Pennsylvania State University was supported by the US National Science Foundation.

Correspondence and requests for materials should be addressed to R.J.C. (e-mail: rcava@princeton.edu).

Self-assembly of regular hollow icosahedra in salt-free catanionic solutions

Monique Dubois*, Bruno Demé†, Thaddée Gulik-Krzywicki‡, Jean-Claude Dedieu‡, Claire Vautrin*, Sylvain Désert*, Emile Perez§ & Thomas Zemb*

* Service de Chimie Moléculaire (CEA/Saclay), F-91191 Gif sur Yvette Cedex, France

† Institut Laue-Langevin, BP 156, F-38042 Grenoble Cedex 09, France

‡ Centre de Génétique Moléculaire, CNRS, F-91198 Gif sur Yvette, France

§ IMRCP, CNRS URA470, 118 route de Narbonne, F-31062 Toulouse, France

Self-assembled structures having a regular hollow icosahedral form (such as those observed for proteins of virus capsids) can occur as a result of biomineralization processes¹, but are extremely rare in mineral crystallites². Compact icosahedra made from a boron oxide have been reported³, but equivalent structures made of synthetic organic components such as surfactants have not hitherto been observed. It is, however, well known that lipids, as well as mixtures of anionic and cationic single chain surfactants, can readily form bilayers^{4,5} that can adopt a variety of distinct geometric forms: they can fold into soft vesicles or random bilayers (the so-called sponge phase) or form ordered stacks of flat or undulating membranes⁶. Here we show that in salt-free mixtures of anionic and cationic surfactants, such bilayers can self-assemble into hollow aggregates with a regular icosahedral shape. These aggregates are stabilized by the presence of pores located at the vertices of the icosahedra. The resulting structures have a size of about one micrometre and mass of about 10¹⁰ daltons, making them larger than any known icosahedral protein assembly⁷ or virus capsid⁸. We expect the combination of wall rigidity and holes at vertices of these icosahedral aggregates to be of practical value for controlled drug or DNA release.

Mixtures of anionic and cationic surfactants in water produce so-called 'catanionic' solutions. After ion pairing, counter-ions form excess salt and induce a high conductivity of the samples that screens electrostatic interactions. A particular type of salt-free catanionic formulation is obtained by using only H⁺ and OH⁻ counter-ions, as no excess salt is then formed by mixing the two surfactants⁹. The resulting salt-free (also called true) catanionic systems can be represented in a ternary phase diagram whose two independent variables are the total volume fraction ϕ of surfactants (assuming surfactant density close to 1) and the molar fraction r of anionic surfactant over total surfactant content.

In dilute salt-free catanionic solutions, we have developed a simple method of producing self-assembled icosahedra of micrometre size in large quantities. Dilute solutions (below $\phi = 10^{-3}$) are prepared by adding water to appropriate amounts of the two surfactants, myristic acid ($M_w = 228.38$, Fluka, purity >99%) and cetyltrimethylammonium hydroxide (CTAOH) ($M_w = 301.54$, ethanolic solution from Acros Organics Co.). The monomer concentration of myristic acid is lower than 10⁻⁶ M, that is, quasi-insoluble in water. The critical micellar concentration, CMC, of CTAOH is slightly lower than CTABr which is 0.0013 mol l⁻¹. Rules for CMC of surfactant mixtures ensure that monomeric concentration is less than 10⁻⁶ M ($\phi < 10^{-4}$)^{10,11}.

The solution is first heated to 60 °C for a few minutes, and then cooled to room temperature with permanent shaking during cooling, until transformation into a slightly bluish clear solution occurs. The solution does not contain any trace of insoluble material. The micrometre-sized icosahedra are formed upon cooling the salt-free catanionic solution, when the anionic component is in excess.

Acceleration of the phase separation between a condensed lamellar phase and a fluid supernatant consisting of icosahedra can be induced by slight centrifugation at 3,000 g. Heating ensures redissolution and aggregates can be restored by cooling down again. Slow dissolution and gentle temperature cycles do not involve significant input of energy other than to disperse the small colloidal objects. Thus, the icosahedral shape is only due to molecular interactions. As in the case of lipid microtubules¹², the aggregates are formed by cooling the sample through a first-order phase transition¹³. In our case, chain crystallization of ion pairs occurs while crossing the chain melting transition, determined by differential scanning calorimetry (DSC) to be in the range 50–65 °C depending on molar fraction r . Dilute solutions (less than 1 g l⁻¹) of icosahedra are metastable: mass measurements and freeze-fracture experiments can be done several weeks after preparation without noticeable change. Addition of glycerol prolongs the stability of the dispersion. However, after a period of the order of several months, a phase separation occurs: the top of the tube contains a thin layer of swollen lamellar phase (L_β).

The region where only icosahedra are formed is found in the range $0.5 < r < 0.75$ and for a total surfactant volume fraction $\phi < 10^{-3}$. The maximum volume fraction where micrometre-size aggregates are in a condition of close packing is $\phi = 3 \times 10^{-3}$, as can be derived from simple geometrical evaluation. The structural electric charge of the aggregates is negative owing to the excess of myristic acid and rises with r up to 32 $\mu\text{C cm}^{-2}$ for a known area per chain of 0.25 nm².

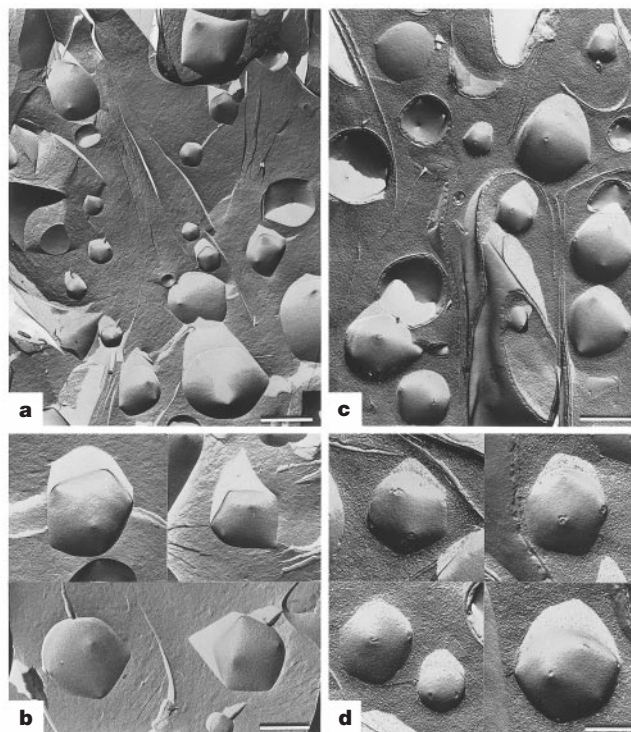


Figure 1 Electron microscopy images of icosahedral aggregates. Freeze fracture (**a, b**) and freeze-etching (**c, d**) obtained for a myristic acid–CTAOH sample ($r = 0.57$, initial weight fraction 0.0163 in a 1:2 glycerol/water solution). Samples were deposited on a thin copper holder, and rapidly quenched in liquid propane. Frozen samples were fractured in vacuum (10⁻⁷ torr) at –125 °C with a liquid-nitrogen-cooled knife in a Balzers 301 freeze-etching unit. The replication was done using unidirectional shadowing with platinum-carbon (35°, thickness 1–1.5 nm). **a, b**, The pores observable at vertices on freeze-fractured samples can be enlarged after the 5-min etching at –105 °C **c, d**. The scale bars represent 1 μm (**a, c**) and 500 nm (**b, d**).

Detection and characterization of dispersions containing icosahedra (Fig. 1) were primarily made by freeze-fracture and freeze-etching electron microscopy. Averaging over several hundreds of imaged objects, we find a narrow radius distribution in the range 700–1,000 nm, corresponding to an edge length of 1 μm . From simple geometrical considerations, and assuming that all objects are single-walled frozen bilayers, we evaluate that 2×10^7 catanionic pairs form one aggregate, which therefore has a molar mass close to 10^{10} Da. As can be seen in Fig. 2, the contour shape observed in freeze-fracture electron microscopy differs from symmetry apparent in cryo-TEM (transmission electron microscopy). At high magnification (Fig. 1b and d), a crucial feature appears: vertices of icosahedra consist of pores. The presence of holes and the associated electrostatic repulsions between them lead to regular polyhedra, which are different from closed-faceted vesicles¹⁴ obtained, for instance, using a strongly adsorbed polymer¹⁵.

For catanionic solutions in an excess of cationic component, nanodisks are formed by rejection of excess charges towards the edges¹⁶. The only other topology to evacuate excess charges is to segregate them in pores¹⁷. Vesicles present at high temperatures⁵ with excess anionic component do not produce nanodisks upon cooling. The presence of excess insoluble fatty acid produces pores at vertices. These can be seen by electron microscopy or confocal microscopy. There are two ways of understanding the formation of pores by cooling vesicles: if the monolayer spontaneous curvature is close to zero while the gaussian curvature is negative, the presence of pores minimizes the bending energy^{18,19}. Producing pores in the bilayer is a way for the system to self-assemble by minimising its curvature energy, since negative gaussian curvature is favoured at high surface charge^{20,21}. Another mechanism at the origin of pore formation would be related to the absence of curvature singularity at vertices where a pore is present. The most frequent morphology of the aggregates is the icosahedron. However, we have observed the formation of other polyhedral shapes. These correspond to the family of structures with less than 12 vertices per polyhedra reviewed by Nelson and Spaepen²⁰. Pores formed in these defect-stabilized structures are a privileged location for cationic solutes, impurities and probably oligopeptides (see Supplementary Infor-

mation). As for catanionic nanodisks, in the absence of salt, molecular segregation is energetically possible owing to overwhelmingly strong electrostatic interactions²¹.

Combined small-angle neutron and light scattering experiments (SANS and SALS, Fig. 3) provide independent proof that closed unilamellar objects of well-defined size are formed. The exact decay of the scattering as the second power of the scattering vector (q) and over an exceptionally wide range of vectors covered by light and neutron scattering at any concentration is evidence of the wall rigidity; the walls do not undulate under brownian motion. As for carbon nanotubes, the persistence length is in the micrometre range. Thus the elastic Young's modulus of the walls is of the order of at least 100 MPa, as for the nanodisks previously described⁹. The size and mass of the aggregates derived from these experiments coincide with those found by freeze-fracture electron microscopy. The minimum located at $q = 0.157 \text{ \AA}^{-1}$ corresponds to the first minimum of the wall form factor oscillation and yields a bilayer thickness of 40 $\text{\AA}$, close to the thickness of the nanodisks made of the same surfactant mixture.

In the dilute regime of salt-free catanionic mixtures, unilamellar vesicles form above the chain melting temperature⁵. During cooling down, the chain crystallization produces a two-dimensional array of alternated cationic and anionic heads. The array of crystallized chains has been characterized by wide-angle neutron scattering, using a mixture of perdeuterated anionic and protonated cationic components. Wide-angle neutron-scattering patterns obtained in the concentrated regime (L_β lamellar phase), suggest the presence of in-plane superstructures formed of alternated chains^{22,23} $-\text{CD}_2-$ and $-\text{CH}_2-$ (see Supplementary Information). However, perfect order does not propagate over large distances, as is clear from the width of the in-plane superstructure peak and the absence of higher-order Bragg reflections. The excess of anionic surfactant is segregated during crystallization of the faces. The amount of excess anionic component is such that 12 pores per vesicle are formed (see the confocal microscopy images indicating the specific location of the charged dye in the Supplementary Information).

Pores formed during crystallization are charged and repel due to the absence of screening salt (Debye screening length of the order of

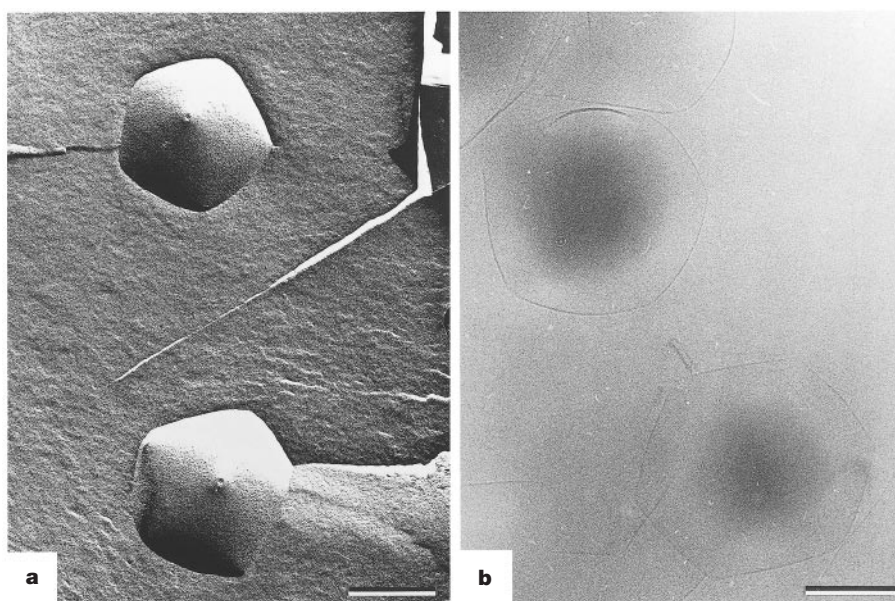


Figure 2 Comparison of freeze-fracture (a) and cryo-TEM (b) images for two adjacent aggregates. We note the systematic presence of five vertices as first neighbours centred on a given vertex. Freeze-fracture suggests pentagonal symmetry around a vertex, whereas the TEM images enhance bilayers normal to the image showing the hexagonal

shape, which is expected for an equatorial fracture through an icosahedron. The icosahedron is the only structure which appears to be hexagonal on projection and pentagonal after fracture. The scale bars represent 250 nm.

100 nm). The structure that minimizes the bending energy with the electrostatic constraints is a regular icosahedron. Any other regular polyhedra would increase the free energy of the aggregate.

Formation of icosahedra with a preferred diameter of 2 μm is controlled both by thermodynamics and kinetics. At high temperature, thermodynamically stable vesicles are present above the chain-melting temperature. During crystallization, bilayers expel laterally a small part of the component in excess, which is not miscible in the two-dimensional crystal. This step is kinetically controlled and leads to the formation of a certain density of pores. If the density of pores is too low or too high, icosahedra cannot form and planar fragments are found (see phase diagram and micrographs in the Supplementary Information).

The method described here is an example of a general principle of formation of icosahedra using given number of surfactant pairs with appropriate chain length and headgroup volumes. The three conditions for formation of regular icosahedra are:

(1) A domain of compositions in the phase diagram where vesicles are the stable form at high temperature. The corresponding region at room temperature is a tie line between a swollen lamellar phase with frozen chains and water.

(2) The excess of surfactant must be insoluble in water and in the frozen bilayer forming the planar rigid walls. This excess surfactant easily forms pores on its own owing to the packing properties of unscreened charged surfactant headgroups²¹.

(3) The part of excess component segregated by the two-dimensional crystal of frozen chains must be available in a sufficient quantity to form twelve pores per vesicle.

These rules apply for all surfactant pairs studied so far. The absence of condition (3) leads to the formation of nanodisks or large

opened crystalline bilayers including pores (see Supplementary Information). Negatively charged icosahedra of identical size are formed in pure water or in water/glycerol mixtures (2:1 by weight).

Typical conductivity of solutions containing icosahedra ($r \approx 0.65$) is $10 \mu\text{S cm}^{-1}$ for an apparent pH of 5.6 measured with a calomel electrode. Strong electrostatic repulsions between icosahedra explain the observed kinetic stability. After a few months at rest, a thin layer of lamellar phase appears at the top of the test tube. The high rigidity of the faces quenches the so-called 'Helfrich' repulsive interaction, when an icosahedron comes into contact with a rigid interface, by at least a factor of hundred. These are very different from soft single-layer or multilayer vesicle²⁴ structures widely used for drug delivery or penetration of creams through the skin in cosmetic applications. Holes at vertices can be covered by an adequate polyelectrolyte, as already described for the preparation of microcapsules²³. This property would be important for controlled drug or DNA release, if the aggregates described here were resistant to the addition of even small quantities of salt. However, phase separation between a lamellar phase (L_β) and nearly pure water is observed at high ionic strength. In the presence of small quantities (millimoles) of salt or impurities, or after incubation in pure water, we have observed, using confocal microscopy, large flocculates made of several icosahedra adhering to each other much as do a bunch of grapes. These globular aggregates, whose typical size is five to ten micrometres, visible in optical microscopy, are built with irregular polyhedra with flat faces. Their morphology is different from the well known myelinic figures, but they show a topology very similar to biliquid foams made from multivesicular liposomes which have been described by Spector *et al.*²⁴.

Thus we have demonstrated that single-wall icosahedra are formed by cooling solutions of vesicles of appropriate composition

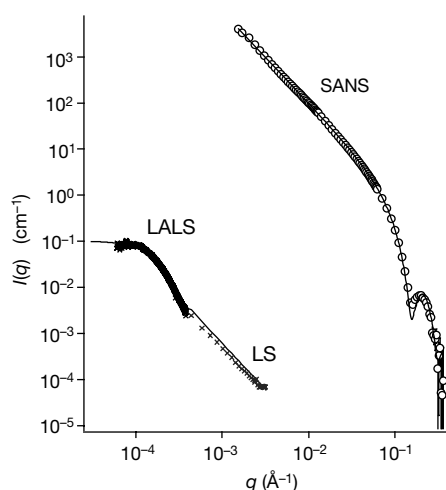


Figure 3 Scattering by dilute solutions of icosahedra. We used small-angle neutron scattering (SANS), low-angle light scattering (LALS) and classical light scattering (LS) covering a q -range: $5 \times 10^{-5} < q < 4 \times 10^{-1} \text{ \AA}^{-1}$. Neutron scattering composition of the sample: $r = 0.636$, $\phi = 2.6 \times 10^{-3}$. Light scattering $r = 0.626$, $\phi = 1.3 \times 10^{-4}$ in glycerol/water. The SANS instrument used D22 (ILL, Grenoble). LS results are compared to a simulation of the scattering produced by independent icosahedra (form factor calculated for a gaussian distribution with average radius of 0.9 μm , and full-width at half-maximum, FWHM = 0.6 μm). SANS shows an ideal q^{-2} dependence over four decades in intensity, followed by the characteristic oscillation of the shell form factor with the first minima at $q = 0.157 \text{ \AA}^{-1}$ corresponding to the membrane thickness. The solid line is a fit to the data using a three-layer model for infinitely flat membranes convoluted by the instrumental resolution function accounting for the wavelength spread and finite spatial resolution of the camera. It yields a bilayer thickness $2t = 40 \text{ \AA}$. The average mass M of the icosahedra is extracted from the LALS data at zero angle, yielding $2 \times 10^{10} \text{ Da/icosahedron}$. ($M = 8.66\rho l^2 2t$, where l is the edge length, and ρ the density close to 1).

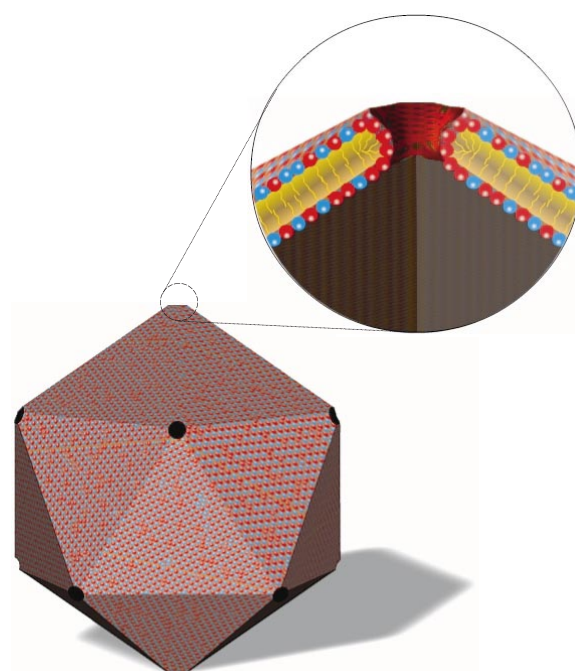


Figure 4 Sketch of the aggregate structure. Frozen alternated hydrocarbon chains produce large flat faces made of about 10^6 ion pairs. These triangular faces, with a typical area of $0.3 \mu\text{m}^2$, result from the juxtaposition of ion pairs tiling an hexagonal lattice. The aggregates are stabilized by the pores (diameter of the order of $150 \text{ \AA}$) produced by about 200 molecules and due to the partial segregation of the anionic surfactant in excess. The regular icosahedron is the structure that minimizes the bending energy of rigid bilayers, as only 30 edges with a dihedral angle of 42° delimit triangular faces.

and in the absence of salt using only two synthetic surfactants. The microstructure of these complex aggregates is sketched in Fig. 4. Self-assembly of icosahedra, owing to defect-free repetition of strong interactions, is one of the fascinating consequences of the quasi-equivalence principle²⁵: highly symmetric structures can be formed with high efficiency by association of identical subunits in the absence of templates. The self-assembled two-dimensional nearly crystalline shell is similar in shape to the structure initially proposed for viruses by Watson and Crick²⁶. In catanionic icosahedra, subunits are combinations of catanionic pairs. Although in viruses the number of subunits is smaller, the type of aggregates described here should obey a similar topology control mechanism: a locally hexagonal lattice folds into 20 equivalent triangles. The association of anionic and cationic surfactants produces particles similar in shape but larger than viruses. □

Received 14 February; accepted 27 March 2001.

- Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, Cambridge, 1961).
- Kimoto, K. & Nishida, I. The crystal habits of small particles of aluminium and silver prepared by evaporation in clean atmosphere of argon. *J. Appl. Phys.* **16**, 941–948 (1966).
- Hubert, H. H. *et al.* Icosahedral packing of B₁₂ icosahedra in boron suboxide. *Nature* **391**, 376–378 (1998).
- Sackmann, E. & Lipowsky, R. *Handbook of Biological Physics* Vol. 1B (ed. Hoff, A. J.) (North-Holland, Amsterdam, 1995).
- Kaler, E. W., Murthy, A. K., Rodriguez, B. E. & Zasadzinski, J. A. N. Spontaneous vesicle formation in aqueous mixtures of single-tailed surfactants. *Science* **245**, 1371–1374 (1989).
- Dubois, M. & Zemb, T. Swelling limits for bilayer microstructures: the implosion of lamellar structure versus disordered lamellae. *Curr. Opin. Colloid Interf. Sci.* **5**, 27–37 (1997).
- Wals, J. *et al.* Tricord protease exists as an icosahedral supermolecule *in vivo*. *Mol. Cell* **1**, 59–65 (1997).
- Klug, A., Franklin, R. E. & Humphrey-Owen, S. P. F. The crystal structure of Tipula iridescent virus as determined by Bragg reflection of visible light. *Biochem. Biophys. Acta* **32**, 203–219 (1959).
- Dubois, M., Gulik-Krzywicki, Th., Demé, B. & Zemb, T. Rigid organic nanodiscs of controlled size: A catanionic formulation. *C. R. Acad. Sci. Paris II C* **19**, 567–575 (1998).
- Scamehorn, J. F. & Hartwell, J. H. in *Surfactant Science Series* (eds Scamehorn, J. F. & Harwell, J. H.) Vol. 33, Ch. 10 (Marcel Dekker Inc., New York, 1989).
- Graciaa, A., Ghoulam, M. B., Marion, G. & Lachaise, J. Critical concentrations and compositions of mixed micelles of sodium dodecylbenzenesulfonate, tetradecyltrimethylammonium bromide and polyethylenephenols. *J. Phys. Chem.* **93**, 4167–4173 (1989).
- Schnur, J. Lipid tubules: a paradigm for molecularly engineered structures. *Science* **262**, 1669–1675 (1993).
- Thomas, B. N., Safinya, C. R., Plano, R. J. & Clark, N. A. Lipid tubules self-assembly: length dependence on cooling rate through a first-order phase transition. *Science* **227**, 1635–1638 (1995).
- Parente, R. A., Höchli, M. & Lentz, B. R. Morphology and phase behavior of two types of unilamellar vesicles prepared from synthetic phosphatidylcholines studies by freeze-fracture electron microscopy and calorimetry. *Biochim. Biophys. Acta* **812**, 493–502 (1985).
- Regev, O., Marques, E. F. & Khan, A. Polymer-induced structural effects on catanionic vesicles: formation of faceted vesicles, disks, and cross-links. *Langmuir* **15**, 642–645 (1999).
- Zemb, T., Dubois, M., Demé, B., Gulik-Krzywicki, T. Self-assembly of flat nanodiscs in salt-free catanionic surfactant solutions. *Science* **283**, 816–820 (1999).
- Hyde, S. T. Topological transformations mediated by bilayer punctures: from sponge phases to bicontinuous monolayers and reversed sponges. *Colloids Surf. A* **129–130**, 207–225 (1997).
- Fogden, A., Daicic, J. & Kidane, A. Undulating charges in fluid membranes and their bending constants. *J. Phys. (Paris) II* **7**, 229–248 (1997).
- Fogden, A., Daicic, J. & Kidane, A. Bending rigidity of ionic surfactant interfaces with variable charge density: the salt-free case. *Colloids Surf. A* **129–130**, 157–165 (1997).
- Nelson, D. R. & Spaepen, F. Polytetrahedral order in condensed matter. *Solid State Phys.* **42**, 1–90 (1989).
- Dubois, M., Belloni, L., Zemb, T., Demé, B. & Gulik-Krzywicki, T. Formation of rigid nanodiscs: Edge formation and molecular separation. *Prog. Colloid Polym. Sci.* **115**, 238–242 (2000).
- Radtchenko, I. L. *et al.* Assembly of alternated multivalent ion/polyelectrolyte layers on colloidal particles. Stability of the multilayers and encapsulation of macromolecules into polyelectrolyte capsules. *J. Colloid Interf. Sci.* **230**, 272–280 (2000).
- Donath, E., Sukhorukov, G. B., Caruso, F., Davis, S. A. & Möhwald, H. Novel hollow polymer shells by colloid-templated assembly of polyelectrolytes. *Angew. Chem. Int. Edn* **37**, 2202–2205 (1998).
- Spector, M. S. & Zasadzinski, J. A. Topology of multivesicular liposomes, a model bicontinuous foam. *Langmuir* **12**, 4704–4708 (1996).
- Klug, A. & Caspar, D. L. D. The structure of small viruses. *Adv. Vir. Res.* **7**, 225–325 (1960).
- Crick, F. H. & Watson, J. D. The structure of small viruses. *Nature* **177**, 473–475 (1956).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank G. Sukhorukov and O. Tiourina for help in confocal optical microscopy, and

I. Erk for cryo-TEM. We also thank J.-L. Sikorav, P. Timmins and B. W. Ninham for critical reading of the manuscript and suggestions.

Correspondence and requests for materials should be addressed to M.D. (e-mail: dubois@scm.saclay.cea.fr).

Stability of atmospheric CO₂ levels across the Triassic/Jurassic boundary

Lawrence H. Tanner*, John F. Hubert†, Brian P. Coffey‡§ & Dennis P. McInerney||

* Department of Geography and Geosciences, Bloomsburg University, Bloomsburg, Pennsylvania 17815, USA

† Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA

‡ Department of Geological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA

|| Environmental Risk Management, Fleet National Bank of Connecticut, Hartford, Connecticut 06102, USA

The Triassic/Jurassic boundary, 208 million years ago, is associated with widespread extinctions in both the marine and terrestrial biota. The cause of these extinctions has been widely attributed to the eruption of flood basalts of the Central Atlantic Magmatic Province^{1–4}. This volcanic event is thought to have released significant amounts of CO₂ into the atmosphere, which could have led to catastrophic greenhouse warming^{5–7}, but the evidence for CO₂-induced extinction remains equivocal. Here we present the carbon isotope compositions of pedogenic calcite from palaeosol formations, spanning a 20-Myr period across the Triassic/Jurassic boundary. Using a standard diffusion model^{8,9}, we interpret these isotopic data to represent a rise in atmospheric CO₂ concentrations of about 250 p.p.m. across the boundary, as compared with previous estimates of a 2,000–4,000 p.p.m. increase^{4,5}. The relative stability of atmospheric CO₂ across this boundary suggests that environmental degradation and extinctions during the Early Jurassic were not caused by volcanic outgassing of CO₂. Other volcanic effects—such as the release of atmospheric aerosols or tectonically driven sea-level change—may have been responsible for this event.

The mass extinction at the end of the Triassic claimed about 80% of all species¹⁰, including most non-dinosaurian archosaurs¹¹. Although extinctions in the terrestrial and marine environments may be slightly asynchronous, they are closely related temporally and undoubtedly share causality¹². This biotic crisis has been attributed previously to bolide impact^{6,13} and sea-level change^{14,15}. However, dating of the best-candidate impact structure (the Manicouagan crater) places the impact roughly 14 Myr earlier¹⁶, and sea-level change fails to explain the fern spike in the terrestrial boundary record⁶. The prevailing view is that this event resulted from the eruptions of the Central Atlantic Magmatic Province (CAMP), which may have produced as much as 7 × 10⁶ km³ of flood basalt in as little as 2 Myr³. CAMP volcanics include the basalts of the Newark Supergroup of eastern North America whose ages of around 200 Myr and stratigraphical proximity to the Triassic/Jurassic boundary are well constrained^{12,17}. A frequently cited deleterious effect of these widespread, massive eruptions is a sudden increase in

§ Present address: ExxonMobil Upstream Research Company, Houston 77252, Texas, USA.

atmospheric CO_2 (p_{CO_2}) from outgassing, resulting in intense global warming^{4–7}.

The estimation of atmospheric palaeo- p_{CO_2} using $\delta^{13}\text{C}$ of pedogenic carbonate and the diffusion-reaction model is now common practice^{8,9,18}. The carbon isotopic composition of soil carbonate is controlled by: the C_4/C_3 vegetation ratio; the depth of carbonate accumulation; the temperature of carbonate precipitation; the isotopic composition of atmospheric carbon, a value well constrained by studies of marine carbonates; and soil productivity ($S(z) = p_{\text{CO}_2,\text{soil}} - p_{\text{CO}_2,\text{atmos}}$), which is largely dependent on climatic regime. Values of palaeo- p_{CO_2} derived from $\delta^{13}\text{C}$ are similar to values from other sources, such as geochemical modelling¹⁹. Elevated palaeo- p_{CO_2} for the Late Triassic has been estimated previously from the isotopic analysis of calcite in palaeosols^{8,9,18}, but reliable data gathered with modern sampling protocols^{9,18} are lacking for the Early Jurassic.

Our data were generated by sampling micritic calcite of pedogenic origin in mature, morphologically similar palaeosols in the rift basins of the Newark Supergroup of eastern North America and a continental basin of the southwestern United States (Fig. 1). The 21 samples comprise a data set that is tightly constrained by field and petrographical criteria to minimize variations in the composition of soil carbonate that result from differences in the depth of calcite accumulation and soil productivity. These samples represent discrete, mudstone-hosted, micritic nodules that occur in mature palaeosol profiles in which the upper part of the profile is characterized by a horizon of coalesced nodules (Bk) or laminar calcite (K). We selected samples from depths of 30 cm or greater, below the top of the Bk or K horizon, thereby minimizing the effects of soil depth^{9,18}. Micritic calcite was selectively removed by microdrilling from thin section billets. The stratigraphic units sampled are: the Chatham Group (Carnian) of the Deep River basin, the New Haven Formation (Norian) of the Hartford basin, the McCoy Brook Formation (Hettangian) of the Fundy basin, and the Owl Rock Formation (Norian) of the Chinle group in the Chinle basin (Fig. 1).

These soils all formed on alluvial floodplains at low palaeolatitudes ($\leq 15^\circ$) and most profiles show features such as gradational lower boundaries, rhizoliths, and circumgranular cracking, which suggest semi-arid to arid conditions, and so have been inferred to represent well-drained soils^{20–23}.

The mean $\delta^{13}\text{C}$ values for calcite in palaeosols from the Chatham Group, New Haven Formation, and Owl Rock Formation coincide within the limits of measurement, averaging -7.4‰ relative to Pee Dee Belemnite (PDB) (Table 1). Calcite from McCoy Brook Formation palaeosols is slightly heavier, with a mean $\delta^{13}\text{C}$ of -7.1‰ PDB (Table 1). This concordance suggests that these palaeosols formed under similar conditions of atmospheric composition and climate (warm, semi-arid to arid). The consistent isotopic compositions of pedogenic calcite is notable because these basins extend across about 17° of palaeolatitude, from the equatorial Deep River basin to the Fundy basin at about 15°N (ref. 24). Application of the diffusion-reaction model to these results requires several assumptions that (admittedly) introduce substantial room for variability in the calculated palaeo- p_{CO_2} . The $\delta^{13}\text{C}_{\text{atmos}}$ is set at -6.5‰ PDB⁹, although variation in the composition of marine carbonates suggests that this value varied slightly during the early Mesozoic¹⁸. The $\delta^{13}\text{C}$ value for organic matter was taken as -24‰ PDB, consistent with direct measurements⁹ and proxy measurements¹⁸ for Upper Triassic and Lower Jurassic palaeosols. Fractionation of soil CO_2 by 1‰ relative to organic matter is assumed⁹. The temperature of carbonate precipitation within the soil is set at 25°C for these low-latitude palaeosols.

Potentially, the greatest source of variation in this calculation is the soil CO_2 productivity $S(z)$, controlled largely by precipitation and drainage characteristics of the soil. The morphologies of these palaeosols suggest formation in well-drained arid to semi-arid soils in which $S(z)$ may vary from 3,000–7,000 p.p.m. (ref. 18). A mean $S(z)$ value of 5,000 p.p.m. was chosen for these calculations. The combined mean $\delta^{13}\text{C}$ of -7.4‰ for Late Triassic pedogenic calcites (Chatham Group, New Haven Formation and Owl Rock Forma-

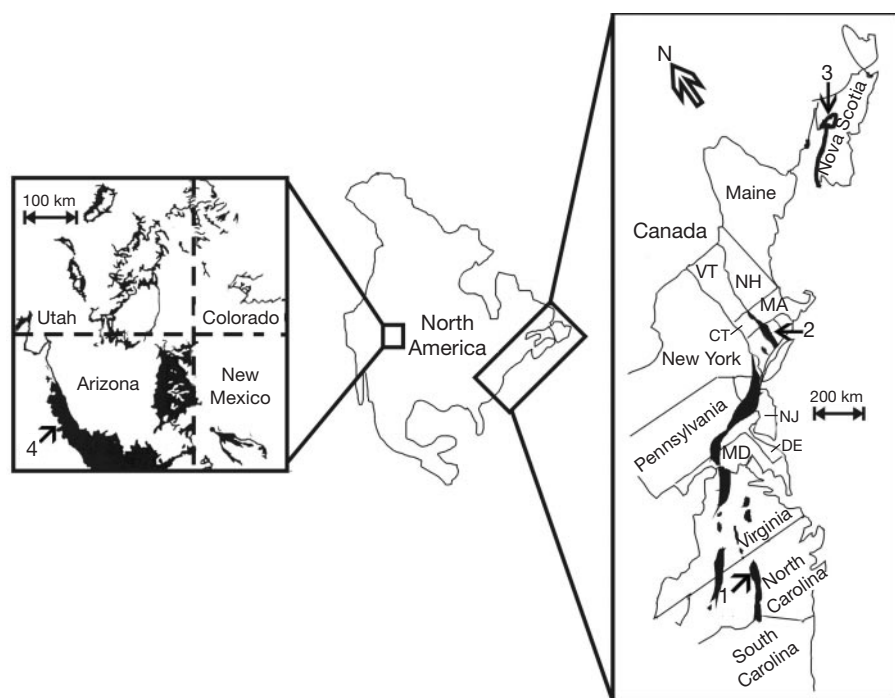


Figure 1 Location of basins containing the sampled palaeosols. Right, exposed basins of the Newark Supergroup. Sampled basins: 1, Durham sub-basin of the Deep River basin (Chatham Group samples from the eastern Durham sub-basin); 2, Hartford basin (New

Haven Formation); and 3, Fundy basin (McCoy Brook Formation). Left, part of the Chinle basin with the outcrop pattern of the Chinle Group shown in black. Location of the sample area of the Owl Rock Formation is indicated (4).

Table 1 Isotopic data

	Sample	$\delta^{13}\text{C}_{\text{‰ PDB}}$
Upper Triassic palaeosols Chatham Group (Carnian)	94-7-1	-7.4
	94-7-2	-7.3
	94-7-2B	-7.4
	94-7-7	-7.4
	94-7-8	-7.3
New Haven Formation (Norian)	S-35	-7.5
	M17-S12	-7.3
	OR-1	-7.8
Owl Rock Formation (Norian)	EC-1	-7.6
	EC-1B	-6.8
	EC-2	-7.4
	EC-2B	-6.8
	EC-3	-7.3
	EC-3B	-7.6
	EC-4	-7.9
Mean	(n = 15)	-7.4 ($\sigma = 0.3$)
Lower Jurassic palaeosols McCoy Brook Formation (Hettangian)	MP1-1	-6.2
	MP1-2	-7.1
	MP1-2B	-7.0
	MP2-2	-7.6
	MP2-2B	-7.5
	MP2-3	-7.3
Mean	(n = 6)	-7.1 ($\sigma = 0.5$)

Analyses performed by Coastal Science Laboratories, Austin, Texas. Values are accurate to 0.2‰.

tion) yields a palaeo- p_{CO_2} value of 2,250 p.p.m., about eight times the modern (pre-industrial) level, with a potential range of 1,350–3,150 p.p.m. (for $S(z)$ values of 3,000 and 7,000). Calculation of palaeo- p_{CO_2} using the McCoy Brook Formation data, yields slightly higher values of 2,480 p.p.m., within a possible range of 1,490–3,470 p.p.m.. Owing to the temperature control on isotope fractionation⁸, the difference between the Upper Triassic and Lower Jurassic palaeosols is eliminated if the mean temperature of pedogenic calcite precipitation were 2 °C lower in the more northerly Fundy basin. Our results are similar to previous estimates of palaeo- p_{CO_2} for the Late Triassic⁸ and the Late Jurassic²⁵; however, these earlier estimates lacked the stratigraphical resolution necessary to examine changes in palaeo- p_{CO_2} across the Triassic/Jurassic boundary. These estimates and ours are contained within the range of error for average palaeo- p_{CO_2} for the latest Triassic and earliest Jurassic calculated by carbon-cycle modelling of 3–8 times the modern level¹⁹. Most significantly, our results differ substantially from previously published analyses for the McCoy Brook Formation (Lower Jurassic), which were obtained by bulk analyses of samples collected without the constraint of position within the palaeosol profile²⁶.

The isotopic composition of oolitic goethite has been cited as evidence for greatly elevated (possibly 18 times the modern level) p_{CO_2} during the Early Jurassic⁴; however, the age of these data are not well constrained and are still subject to interpretation. Within the accuracy of measurement, our data do not indicate a significant increase in palaeo- p_{CO_2} across the Triassic/Jurassic boundary. We do not therefore support an interpretation of sudden global warming resulting from CO_2 outgassing during CAMP volcanism. Suggestions of large increases in Early Jurassic palaeo- p_{CO_2} due to CAMP eruptions seem to be made on the basis of gross overestimates of the CO_2 contribution from this magmatism. Recent estimates of the maximum volume of CAMP basalt³ are comparable to the largest estimates for the Deccan Traps flood basalts²⁷, the eruptions of which are estimated to have increased atmospheric CO_2 by only 200–250 p.p.m.v. (against a Late Cretaceous background of about 600 p.p.m.v.)²⁸. This calculation, moreover, is based on an estimate of 0.5% CO_2 in the basaltic magma that may be overgenerous²⁹. Therefore, we conclude that eruptions of the CAMP flood basalts were unlikely to have greatly increased palaeo- p_{CO_2} during the Early Jurassic. Alternative mechanisms for extinction need to be explored more fully, particularly short-term

cooling or acidification of the atmosphere and oceans from SO_2 aerosols^{6,30}, and rapid sea-level change driven by plume-driven thermal doming and collapse⁷. Although little evidence has been presented to support the former¹⁴, the latter hypothesis has the advantage of explaining the transgressive–regressive couplet that spans the Triassic/Jurassic boundary^{14,15} and the widespread record of Early Jurassic marine anoxia¹⁴. □

Received 16 January; accepted 2 April 2001.

- Stothers, R. B. Flood basalts and extinction events. *Geophys. Res. Lett.* **20**, 1399–1402 (1993).
- Courtillot, V. Mass extinctions in the last 300 million years: one impact and seven flood basalts? *Israel J. Earth Sci.* **43**, 255–266 (1994).
- Marzoli, A. et al. Extensive 200-million-year-old continental flood basalts of the central Atlantic province. *Science* **284**, 616–618 (1999).
- Yapp, C. J. & Poeths, H. Carbon isotopes in continental weathering environments and variations in ancient atmospheric CO_2 pressure. *Earth Planet. Sci. Lett.* **137**, 71–82 (1996).
- McElwain, J. C., Beerling, D. J. & Woodward, F. I. Fossil plants and global warming at the Triassic–Jurassic boundary. *Science* **285**, 1386–1390 (1999).
- Olsen, P. E. Giant lava flows, mass extinctions, and mantle plumes. *Science* **284**, 604–605 (1999).
- Hallam, A. The end-Triassic extinction in relation to a superplume event. *Geol. Soc. Am. Abs. Prog.* **37**, 380 (2000).
- Cerling, T. E. Carbon dioxide in the atmosphere: evidence from Cenozoic and Mesozoic paleosols. *Am. J. Sci.* **291**, 377–400 (1991).
- Cerling, T. E. Stable carbon isotopes in paleosol carbonates. *Spec. Publ. Int. Ass. Sediment.* **27**, 43–60 (1999).
- Sepkoski, J. J. Jr. in *Global Events and Event Stratigraphy in the Phanerozoic* (ed. Walliser, O. H.) 35–51 (Springer, Berlin, 1996).
- Curtis, K. & Padian, K. An Early Jurassic microvertebrate fauna from the Kayenta Formation of northeastern Arizona; microfaunal change across the Triassic–Jurassic boundary. *Paleobios* **19**, 19–37 (1999).
- Pálffy et al. Timing the end-Triassic mass extinction: first on land, then in the sea? *Geology* **28**, 39–42 (2000).
- Olsen, P. E., Shubin, N. H. & Anders, M. H. New Early Jurassic tetrapod assemblages constrain Triassic–Jurassic tetrapod extinction event. *Science* **237**, 1025–1029 (1987).
- Hallam, A. & Wignall, P. B. *Mass Extinctions and Their Aftermath* (Oxford, New York, 1997).
- Hallam, A. & Wignall, P. B. Mass extinctions and sea-level changes. *Earth Sci. Rev.* **48**, 217–250 (1999).
- Hodych, J. P. & Dunning, G. R. Did the Manicouagan impact trigger end-of-Triassic mass extinction? *Geology* **20**, 51–54 (1992).
- Fowell, S. J. & Olsen, P. E. Time calibration of Triassic/Jurassic microfossil turnover, eastern North America. *Tectonophysics* **222**, 361–369 (1993).
- Ekart, D. D., Cerling, T. E., Montañez, I. P. & Tabor, N. J. A 400 million year carbon isotope record of pedogenic carbonate: implications for paleoatmospheric carbon dioxide. *Am. J. Sci.* **299**, 805–827 (1999).
- Berner, R. A. Enriched: GEOCARB II; a revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **294**, 56–91 (1994).
- Hubert, J. F. Paleosol caliche in the New Haven Arkose, Newark Group, Connecticut. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **24**, 151–168 (1978).
- Coffey, B. P. & Textoris, D. A. Paleosols and paleoclimatic evolution, Durham sub-basin, North Carolina. *Geol. Nat. Hist. Sur. Connecticut Misc. Rep.* **1**, 6 (1996).
- Tanner, L. H. in *The Continental Jurassic* (ed. Morales, M.) 656–574 (Museum of Northern Arizona, Flagstaff, Arizona, 1996).
- Tanner, L. H. Palustrine-lacustrine and alluvial facies of the (Norian) Owl Rock Formation (Chinle Group), Four Corners region, southwestern U.S.A.: implications for Late Triassic paleoclimate. *J. Sedim. Res.* **70**, 1280–1289 (2000).
- Olsen, P. E. Stratigraphic record of the early Mesozoic breakup of Pangea in the Laurasia–Gondwana rift system. *Annu. Rev. Earth Planet. Sci.* **25**, 337–401 (1997).
- Ekart, D. D. & Cerling, T. E. PCO_2 during deposition of the Late Jurassic Morrison Formation and other paleoclimatic/ecologic data as inferred by stable carbon isotope analyses. *Geol. Soc. Am. Abs. Prog.* **28**, 252 (1996).
- Suchecki, R. K., Hubert, J. F. & De Wet, C. B. Isotopic imprint of climate and hydrogeochemistry on terrestrial strata of the Triassic–Jurassic Hartford and Fundy rift basins. *J. Sed. Res.* **58**, 801–811 (1988).
- Coffin, M. F. & Eldholm, O. Scratching the surface: estimating dimensions of large igneous provinces. *Geology* **21**, 515–518 (1993).
- Caldeira, K. G. & Rampino, M. R. Deccan volcanism, greenhouse warming, and the Cretaceous/Tertiary boundary. *Geol. Soc. Am. Spec. Pap.* **247**, 117–123 (1990).
- Wallace, P. & Anderson, A. T. in *Encyclopedia of Volcanoes* (ed. Sigurdsson, H.) 149–170 (Academic, New York, 2000).
- Parfitt, E. A. & Wilson, L. Impact of basaltic eruptions on climate. *Geol. Soc. Am. Abs. Prog.* **32**, 501 (2000).

Acknowledgements

This study was supported by grants from the Research and Disciplinary Fund and the Special Initiative Fund, Bloomsburg University (L.H.T.), and from the Connecticut Department of Environmental Protection (D.P.I.). We thank the Navajo nation for permitting us to conduct research on reservation lands. We also thank S. Driese, D. Fastovsky and A. Hallam for their comments on this manuscript.

Correspondence and requests for materials should be addressed to L.H.T. (e-mail: lhtann@sunlink.net).

A phenomenological model for precursor volcanic eruptions

Thierry Menand* & Stephen R. Tait*

Laboratoire de Dynamique des Systèmes Géologiques, Institut de Physique du Globe de Paris, 4, place Jussieu, 75252 Paris cedex 05, France

Intense explosions of relatively short duration frequently precede large explosive and effusive volcanic eruptions—by as much as weeks to months in the case of very viscous magmas^{1–6}. In some cases, such pre-eruption activity has served as a sufficient warning to those living in the vicinity to evacuate and avoid calamity¹. Precursor events seem to be related to the formation of a magma pathway to the surface, but their precise interpretation is a long-standing puzzle. It has been inferred from theoretical studies that exsolution of volatiles might create an almost separate gas pocket at the tip of a propagating dyke^{7–9}. Here we explain the role that such a process may have, using a laboratory study of the transient propagation of a liquid-filled crack with a gas pocket at its tip that grows with time. We show that once the gas pocket acquires sufficient buoyancy to overcome the fracture resistance of the host solid the dynamics of the gas pocket, rather than those of the liquid, determine the velocity of the crack tip. Furthermore, we find that the gas can ultimately separate from the liquid. We propose that fast-moving, gas-rich pockets reaching the surface ahead of the main liquid-filled fissure could be the origin of many precursor eruptions.

Precursory activity associated with a number of carefully monitored volcanic eruptions seems to fit into a general pattern. Quasi-continuous volcanogenic seismic tremor associated with sporadic explosions and gas emissions at the surface are observed during several weeks before the arrival of the main magma conduit. Explosions are intense but of short duration, and contain variable amounts (sometimes only very little) of juvenile magmatic component. This pattern occurs whether the eruption is ultimately an explosive plinian event^{3–5,10–12} or dome-forming^{6,13,14}. This activity may sometimes not be followed by a large eruption. Although the term ‘precursory’ seems inappropriate in such cases, this activity may well be part of the same fundamental pattern, but the main magma conduit stalls and never reaches the surface, remaining as a shallow intrusion. At Usu volcano, Japan, months of degassing from craters followed initial intense explosions on 31 March 2000, but at present the main magma conduit has not reached the surface, and seems unlikely now to do so. This pattern follows that of other eruptions at Usu during the last century, for example in 1943, when the lava dome Showa-Shinzan formed¹⁵. Precursory activity is thus a common prelude to large volcanic events. Here we provide a framework in which precursory activity can be interpreted. This can help ultimately to gain understanding of its expected timescale at a volcano of a given type, which is of obvious practical use for hazard assessment.

As silicate magmas ascend towards the Earth’s surface, the reduction of pressure causes dissolved volatile components to exsolve and form gas bubbles. At the tip of a vertically oriented fissure, pressure is lowest because there is least rock overburden. In addition, the high viscosities of magmatic liquids implies large head losses as the fissure tapers to vanishing thickness^{7,8}. As the fissure propagates towards the surface, ambient pressure decreases, driving further exsolution and expansion of the volatiles. Moreover, because

of the no-slip boundary condition at the dyke wall, magma near the dyke wall moves more slowly than at the tip and, from mass balance, magma near the dyke centre moves faster than the tip. This means that fresh magma that has not been degassed is always being transported into the low-pressure region, where it should turn into foam. The subsequent lowering of magma density provides a buoyancy force that tends to keep it there. Intense shearing in this region may cause the foam to break down and allow the gas to collect into a separate pocket at the tip, which grows during propagation. The complex dynamics of this process go beyond the scope of our work; however, we infer that bubbles do not need to move significant distances with respect to the surrounding magma to feed the gas pocket. One theoretical investigation has shown, in the framework of a steadily propagating fissure with a fixed volume of gas at the tip, that the fissure tends to pinch off between the gas and the following liquid⁸. We performed laboratory experiments to look at the transient propagation of a liquid-filled crack with a gas pocket at its tip that grows with time.

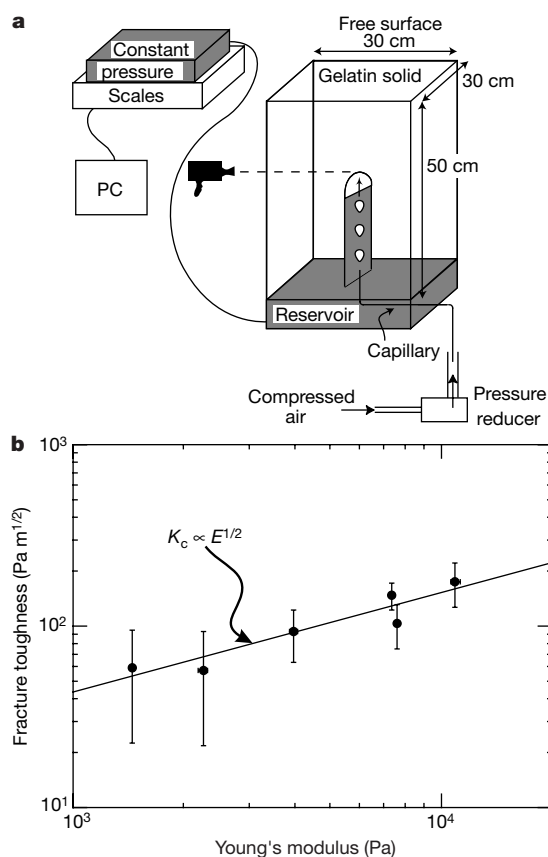


Figure 1 Experimental apparatus and rheology of gelatin. **a**, Gelatin was solidified in an acrylic tank under hydrostatic conditions. Liquid was then injected through the gelatin from a reservoir under constant overpressure. To simulate exsolution and expansion of volatiles at the crack tip, air was continuously injected into the liquid-filled fissure some time after it began to propagate. The air collected at the fissure tip. Dyed aqueous solutions of four different viscosities were used: 1, 41, 126 and 342 mPa s, achieved by dissolving small amounts (less than 0.5 wt%) of hydroxyethylcellulose. These low polymer concentrations allowed solution viscosities to be varied without affecting other properties. Propagation velocity was recorded by videotaping the experiment; liquid injection rate was monitored on a personal computer by measuring output from the reservoir. Experimental durations were less than 5 min, a time period short enough to consider the rheological behaviour of the gelatin to be purely elastic²⁴. **b**, Relationship between Young's modulus E and fracture toughness K_c of the gelatin. E was calculated from the measured vertical deflection induced by a weight on the free surface of the gelatin. K_c was defined as the stress intensity factor at the tip of a fissure of known geometry and internal pressure distribution when gelatin is on the verge of failure.

* Present address: BP Institute for Multiphase Flow, University of Cambridge, Madingley Rise, Madingley Road, Cambridge CB3 0EZ, U.K. (T.M.); and TH Huxley School of Environment, Earth Science and Engineering, Imperial College of Science, Technology and Medicine, RSM Building, Prince Consort Road, London SW7 2BP, U.K. (S.R.T.).

$$U \propto h_b^2 \quad (1)$$

Experimental apparatus and techniques are described in Fig. 1a. Young's modulus and fracture toughness have been measured independently *in situ* in our experimental apparatus. These conform to the expected theoretical relationship for a purely elastic solid^{16,17} (Fig. 1b). This knowledge of the fracture toughness of the gelatin provides an important technical basis for the study. Figure 2 shows pictures of two different stages of an experiment. Initially the liquid-filled crack propagates vertically and radially owing to the reservoir overpressure (Fig. 2a). After we began to inject air into the fissure, a bubble formed at its tip. When it became higher than about 2 cm, the shape of the crack was deformed markedly in both vertical cross-sections (Fig. 2b, c). Moreover, the crack velocity (which is the bubble velocity) increased abruptly while the liquid injection rate stayed constant (Fig. 3), indicating that the crack is closing. Having both a closing crack and an accelerating air bubble indicates that the bubble may separate from the liquid. Liquid injection rate was a decreasing function of liquid viscosity (Fig. 3b).

The controlling influence of the air pocket on fissure propagation once it reached a critical height h_c is better seen on a graph of crack velocity as a function of the height of the air bubble h_b for heights greater than h_c (Fig. 4). Whatever the viscosity of the fluid and the injection rate of air, the crack velocity U is proportional to the square of the bubble height:

$$h_c = \left[\frac{4K_c}{3\rho_g g} \sqrt{\frac{2}{\pi}} \right]^2 \quad (2)$$

where ρ_g is the density of the gelatin solid and g the gravitational acceleration. The gelatin solids in our experiments had a fracture toughness $K_c \approx 54 \text{ Pa m}^{1/2}$, which gives $h_c \approx 3 \text{ cm}$, a value close to the 2 cm suggested by our observations. We conclude that once the bubble buoyancy overcomes the fracture resistance of the gelatin it controls the crack tip velocity.

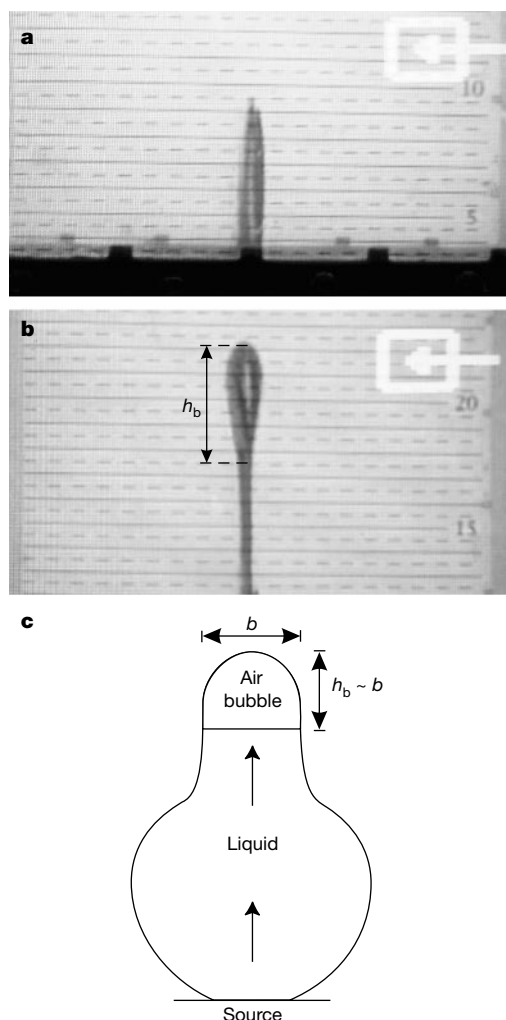


Figure 2 Photographs of an experiment. **a**, Initially the fissure propagates owing to the reservoir overpressure. **b, c**, As the injected air accumulates at the tip it dramatically deforms the crack and dictates its width (**b**) as well as its breadth (**c**).

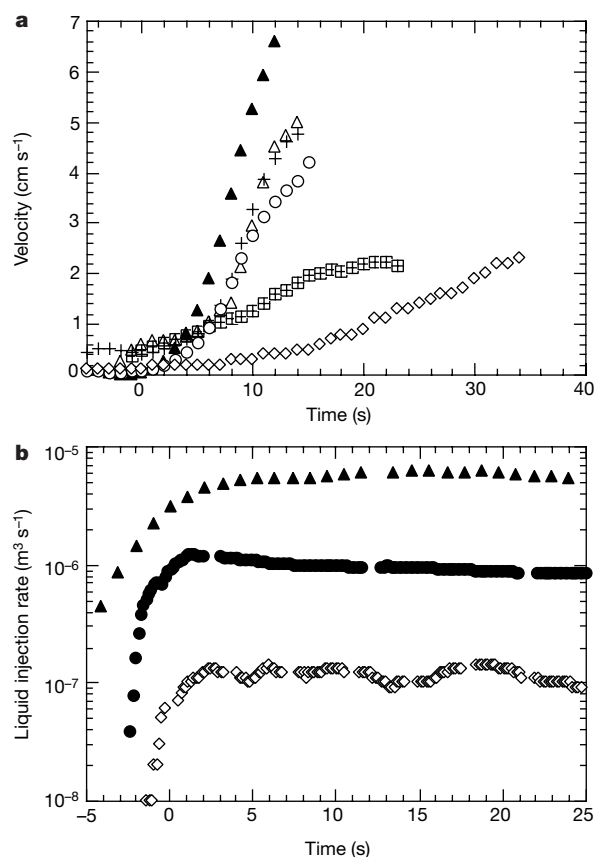


Figure 3 Time evolution of velocity and injection rate during propagation. **a, b**, The velocity (**a**) and the liquid injection rate (**b**) of the crack for different experiments as a function of time. The time origin corresponds to the moment at which the air bubble reached its critical height h_c . Each experiment corresponds to a given fluid viscosity. Diamond, 0.342 Pa s; filled circle, 0.126 Pa s; open circle, 0.041 Pa s; and other symbols, 10^{-3} Pa s.

The velocity of the fissure increases with the volume of the gas pocket; however, the velocity of the liquid is limited by the viscous pressure gradient inside the fissure tail, which balances and is therefore equal to liquid buoyancy. Assuming that the fissure tail has a steady-state horizontal elliptical cross-section and equating the kinematically determined tail width with the Poiseuille flow width, we obtain the fissure velocity u_s above which separation occurs between the gas pocket and the liquid:

$$u_s = \left[\frac{8q^2 \Delta \rho g}{\eta \pi^2 b^2} \right]^{\frac{1}{3}} \approx \left[\frac{8q^2 \Delta \rho g}{\eta \pi^2 h_b^2} \right]^{\frac{1}{3}} \quad (3)$$

where q is the liquid flux, $\Delta \rho$ the difference between the density of the solid and the density of the liquid, η the liquid viscosity and b the breadth of the crack, which is approximately equal to the height of the gas pocket h_b (Fig. 2c). Once fissure velocity exceeds u_s , the gas pocket effectively separates from the liquid-filled fissure; a small quantity of liquid is generally carried along with the gas pocket because the viscosity of the liquid prevents clean separation of the phases.

The velocity at which the gas pocket will propagate once it has separated remains unknown. We have shown that gelatin is behaving as an elastic, brittle solid and that the stress intensity factor is greater than its fracture toughness. Furthermore, we find that crack tip velocity is a simple function of the height of the gas pocket. The fundamental control in this regime is presumably the mechanics of failure and crack propagation at the tip in response to the buoyancy force of the gas. Another possibility is that the velocity of the gas pocket is limited by the propagation of elastic waves in the solid and is comparable to the velocity of such waves. If the host solid is rock, these are of the order of km s^{-1} , considerably greater than those predicted by the regime observed in our experiments. Dyke velocity can be estimated by monitoring micro-seismic events due to cracking at its tip²⁰. Therefore careful analysis of seismicity just before precursor eruptions could potentially resolve the issue. The precise mechanical interpretation, however, is not critical to our proposed model for precursor eruptions. The essential point is that the gas pocket will move faster—often considerably faster—than the main liquid-filled fracture, because liquid viscosity limits the velocity of the latter.

In our experiments, the period of time between the beginning of precursor events and the main eruption, as well as the volume of erupted gas, depends mainly on liquid viscosity; it controls u_s and thus the separation depth. For real magmas, however, the separation depth should also depend on volatile content and hence saturation pressure. The solid is fractured ahead of the liquid and is character-

ized by a lower fracture toughness; new gas pockets would need a lower critical height h_c to separate (equation (2)), and could potentially propagate at greater velocity^{21,22}. New gas pockets forming at the tip of the liquid crack could possibly catch up with the first one. However, if a little magma is transported with the gas phase, it could aid in 'healing' the fractured rock after passage of the gas-rich phase; it seems probable that some magma would be left behind as the walls close behind the main gas pocket and could potentially freeze.

Further research, both observational and theoretical, is required to give more quantitative predictions. Nevertheless, this new framework provides a simple interpretation of precursor eruptions, and can be improved to give more precise estimates of the duration of precursor activity in specific cases. One general prediction is that precursor activity should be shorter in the case of basaltic volcanoes than for silicic ones. This seems to be supported by accounts of the Paricutin eruption, which indicate that gas and fine ash emissions began only hours before the main magma conduit reached the surface²³, a time period considerably shorter than the weeks to months typical of silicic eruptions. □

Received 11 December 2000; accepted 26 March 2001.

- Heiken, G. & McCoy, F. Jr Caldera development during the Minoan eruption, Thira, Cyclades, Greece. *J. Geophys. Res.* **89**, 8441–8462 (1984).
- Gardner, J. E., Carey, S. & Sigurdsson, H. Plinian eruptions at Glacier Peak and Newberry volcanoes, United States: implications for volcanic hazards in the Cascade Range. *Geol. Soc. Am. Bull.* **110**, 173–187 (1998).
- Newhall, C. G. & Punongbayan, R. S. *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines* (Philippine Institute of Volcanology and Seismology, Univ. Washington Press, 1996).
- Foxworthy, B. L. & Hill, M. Volcanic Eruptions of 1980 at Mount St. Helens. The First 100 Days. *US Geol. Survey Prof. Pap.* **1249**, 1–125 (1982).
- Lipman, P. W. & Mullineaux, D. R. The 1980 Eruption of Mount St. Helens. *US Geol. Survey Prof. Pap.* **1250**, 1–844 (1981).
- Young, S. R. et al. Overview of the eruption of the Soufriere Hills volcano Montserrat, 18 July 1995 to December 1997. *Geophys. Res. Lett.* **25**, 3389–3392 (1998).
- Barenblatt, G. I. The mathematical theory of equilibrium cracks in brittle fracture. *Adv. Appl. Mech.* **7**, 55–129 (1962).
- Lister, J. R. Buoyancy-driven fluid fracture: the effects of material toughness and of low-viscosity precursors. *J. Fluid. Mech.* **210**, 263–280 (1990).
- Rubin, A. M. Tensile fracture of rocks at high confining pressure: implications for dyke propagation. *J. Geophys. Res.* **98**, 15919–15935 (1993).
- Daag, A. S. et al. in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines* (eds Newhall, C. G. & Punongbayan, R. S.) 409–414 (Philippine Institute of Volcanology and Seismology, Univ. Washington Press, 1996).
- Harlow, D. H. et al. in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines* (eds Newhall, C. G. & Punongbayan, R. S.) 285–305 (Philippine Institute of Volcanology and Seismology, Univ. Washington Press, 1996).
- White, R. in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines* (eds Newhall, C. G. & Punongbayan, R. S.) 307–327 (Philippine Institute of Volcanology and Seismology, Univ. Washington Press, 1996).
- Aspinall, W. P. et al. Soufriere Hills eruption, Montserrat, 1995–1997: volcanic earthquake locations and fault plane solutions. *Geophys. Res. Lett.* **25**, 3397–3400 (1998).
- Young, S. R. et al. Monitoring SO_2 emission at the Soufriere Hills volcano: implications for changes in eruptive conditions. *Geophys. Res. Lett.* **25**, 3681–3684 (1998).
- Mimatsu, M. *Showa-Shinzan Diary* (Suda-Seihan, Sapporo, Japan, 1995).
- Griffith, A. A. The phenomena of rupture and flow in solids. *Phil. Trans. R. Soc. Lond. A* **221**, 163–198 (1920).
- Irwin, G. R. Analysis of stresses and strains near the end of a crack traversing a plate. *J. Appl. Mech.* **24**, 361–364 (1957).
- Heimpel, A. M. & Olson, P. in *Magmatic Systems* (ed. Ryan, M. P.) 223–240 (Academic Press, San Diego, 1994).
- Rubin, A. M. Dike ascent in partially molten rock. *J. Geophys. Res.* **103**, 20901–20919 (1998).
- Einarsson, P. & Brandsdóttir, B. Seismological evidence for lateral magma intrusion during the July 1978 deflation of the Krafla volcano in NE-Iceland. *J. Geophys.* **47**, 160–165 (1980).
- Takada, A. Experimental study on propagation of liquid-filled crack in gelatin: shape and velocity in hydrostatic stress condition. *J. Geophys. Res.* **95**, 8471–8481 (1990).
- Koyaguchi, T. & Takada, A. An experimental study on the formation of composite intrusions from zoned magma chambers. *J. Volcanol. Geotherm. Res.* **59**, 261–267 (1994).
- Foshag, W. & Gonzalez, J. Birth and development of Paricutin volcano. *U. S. Geol. Surv. Bull. D* **965**, 355–485 (1956).
- Richard, R. Jr & Mark, R. Gelatin models for photoelastic analysis of gravity structures. *Exp. Mech.* **6**, 30–38 (1966).

Acknowledgements

We thank A. M. Rubin for comments and D. L. Sahagian. We also thank A. Agnon and C. Jaupart for discussions and G. Bienfait for help with experiments.

Correspondence and requests for materials should be addressed to T.M. (e-mail: thierry@bpi.cam.ac.uk).

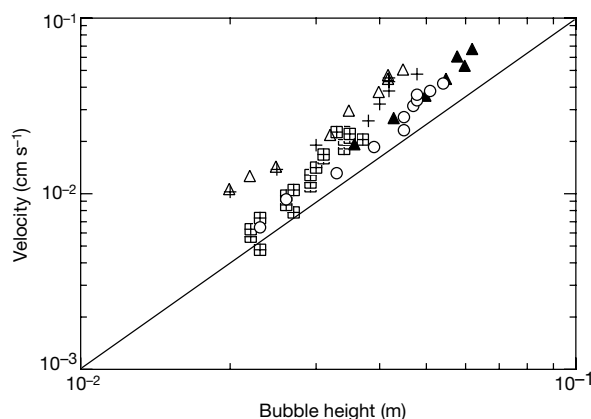


Figure 4 Crack velocity versus gas pocket height. The velocity of the air bubble as a function of its height for injected liquids of two different viscosities (1 mPa s and 41 mPa s) and different injection rates of air for a given fluid. The symbols are as in Fig. 3.

Reykjanes 'V'-shaped ridges originating from a pulsing and dehydrating mantle plume

Garrett Ito

Department of Geology, University of California, Davis, California 95616, USA

Prominent crustal lineations straddle the Reykjanes ridge, south of Iceland (Fig. 1). These giant V-shaped features are thought to record temporal variations in magma production at the Reykjanes ridge axis, associated with along-axis flow of Icelandic plume material¹. It has been proposed that this flow is channelled preferentially along the ridge axis²⁻⁴, and that temporal variability is induced by fluctuations of the Iceland plume itself⁴⁻⁷ or, alternatively, by relocations of the ridge axis on Iceland⁸. Here I present a geodynamic model that predicts the formation of crustal V-shaped ridges from a pulsing and radially flowing mantle plume. In this model, plume pulses produce mantle temperature perturbations that expand away from the plume in all directions beneath the zone of partial melting. The melting zone has a high viscosity owing to mantle dehydration at the onset of partial melting⁹. This high-viscosity region allows for reasonable variations in crustal thickness, produces crustal Vs that extend hundreds of kilometres along the axis, and prevents the plume material from being preferentially channelled along the ridge axis. The angle of the crustal V-shaped features relative to the ridge axis reflects the rate of lateral plume flow, which remains several times greater than the ridge half-spreading rate over the length of a crustal V. Consequently, this radially expanding plume produces lineations in crustal thickness and free-air gravity anomalies that appear to be nearly straight.

This model builds on previous three-dimensional numerical simulations of a ridge-centred plume^{10,11}. Mantle convection is simulated by solving the equations describing conservation of mass, momentum and energy in a fluid of zero Reynolds number and moderately temperature-dependent viscosity¹². An essential feature of this model is an increase in viscosity by a factor of 100 near the base of the primary melt production zone. Such an increase in viscosity is likely to occur as water is extracted from the mantle at the very early stages of decompression partial melting⁹.

The mantle flow and temperature distribution of a steady-state, ridge-centred plume is shown in Fig. 2a. Low-viscosity and thermally buoyant plume material rises beneath the ridge axis, spreads laterally beneath the high-viscosity melting zone, and attains a steady-state, along-axis width^{10,11,13,14}. In the overlying melting zone, high viscosities minimize along-axis and vertical flow, and upwelling occurs primarily to accommodate the spreading of the two plates. This slow, passive upwelling allows a plume of relatively high excess temperature and narrow radius^{15,16} to produce crustal thicknesses¹¹ that are consistent with seismic observations on Iceland (for example, see refs 17 and 18).

I now vary the volume flux of upwelling plume material as a periodic function in time. This is done by varying the imposed radius of the plume stem about the steady-state value of 100 km. The period of variation of 8 Myr is in the range of 5–10 Myr estimated for the age contrasts between adjacent V-shaped ridges in the North Atlantic Ocean^{14,19}. A possible cause for such flux perturbations is the surfacing of solitary waves in the Iceland mantle plume. Solitary waves have been shown to be stable features that initiate from perturbations in ascending plumes of buoyant viscous fluid^{20,21}. The maximum temperature anomaly of the plume stem remains constant.

The flux variations generate pulses in the plume that rise up the stem and expand outward beneath the boundary at which viscosity increases due to dehydration. Figure 2 shows temperature and flow at three time steps of a calculation in which the plume flux varies between $(6.5 \pm 4.9) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$. At a time of 2 Myr after the initiation of pulsing, the plume has reached its maximum radius. Because the stem is wider than the steady-state model, a greater volume of hot material flows more rapidly up the plume stem. At positions near the outer portions of the stem, material is hotter than in the steady-state model because of the greater plume width. Temperatures at the centre of the stem are unchanged. The leading portion of the pulse has already struck the dehydration boundary and has begun to expand away from the plume stem as an annulus of hotter-than-normal material.

At 4 Myr the basal radius of the plume stem has returned to the steady-state value and the pulse is now expanding away from the plume stem beneath the dehydration viscosity boundary (Fig. 2c). Compared to the steady-state condition, the 1,450 and 1,500 °C isotherms extend a greater radial distance from the centre of the plume, indicating hotter material in these regions. Temperature perturbations are greatest (the maximum is +56 °C) at radial distances of ~75–375 km. A cross-section at a depth of 135 km shows that temperature perturbations diminish with distance away from the ridge axis. This is a result of conductive cooling beneath a plate of increasing age.

Finally, at 7 Myr the plume stem is in a contracted form (Fig. 2d). Material surrounding this thinner plume stem is cooler than in the steady-state model and an annulus of cooler-than-normal material is expanding away from the stem beneath the dehydration boundary. The original, high-flux pulse is now 300–500 km away from the plume stem, though the magnitude of temperature perturbations has

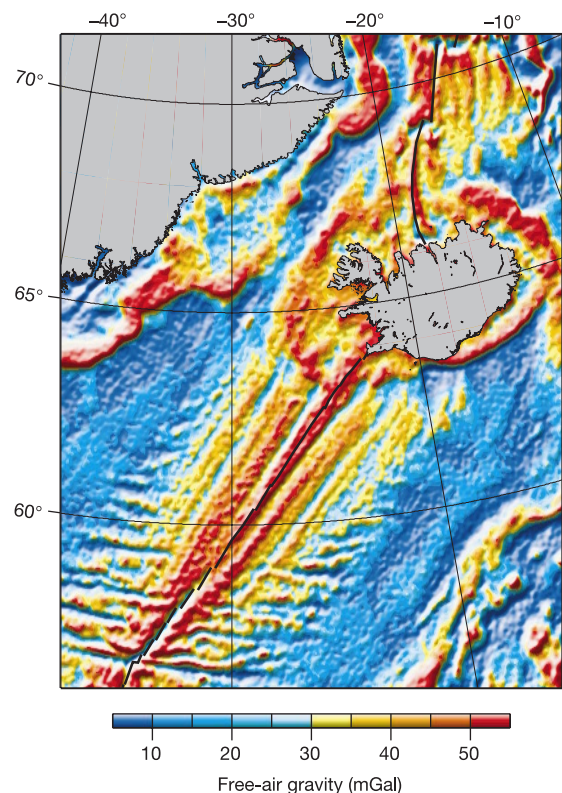


Figure 1 Satellite-derived free-air gravity anomalies in region surrounding Iceland²⁷. Reykjanes ridge is the linear high (bold line) that bisects V-shaped features pointing southward from Iceland. Similar though less regular lineations are also evident near the Kolbeinsey ridge to the north.

diminished (the maximum is $+35^{\circ}\text{C}$). At later times, this pulse will continue to radiate away from the plume stem until its effects are minimal beyond a distance of ~ 600 km. The cooler-than-normal material will follow the leading pulse and then a new high-flux pulse will rise to start a new cycle.

The above dynamics result in alternating high and low tempera-

ture perturbations that propagate along the ridge axis beneath the partial melting zone. This model thus presents a mechanism for generating variations in mantle temperature, as described in ref. 4, and magma production anomalies that propagate along-axis, as first conceptualized in ref. 1. The pattern of crustal thickness predicted by assuming all melts produced in the mantle migrate

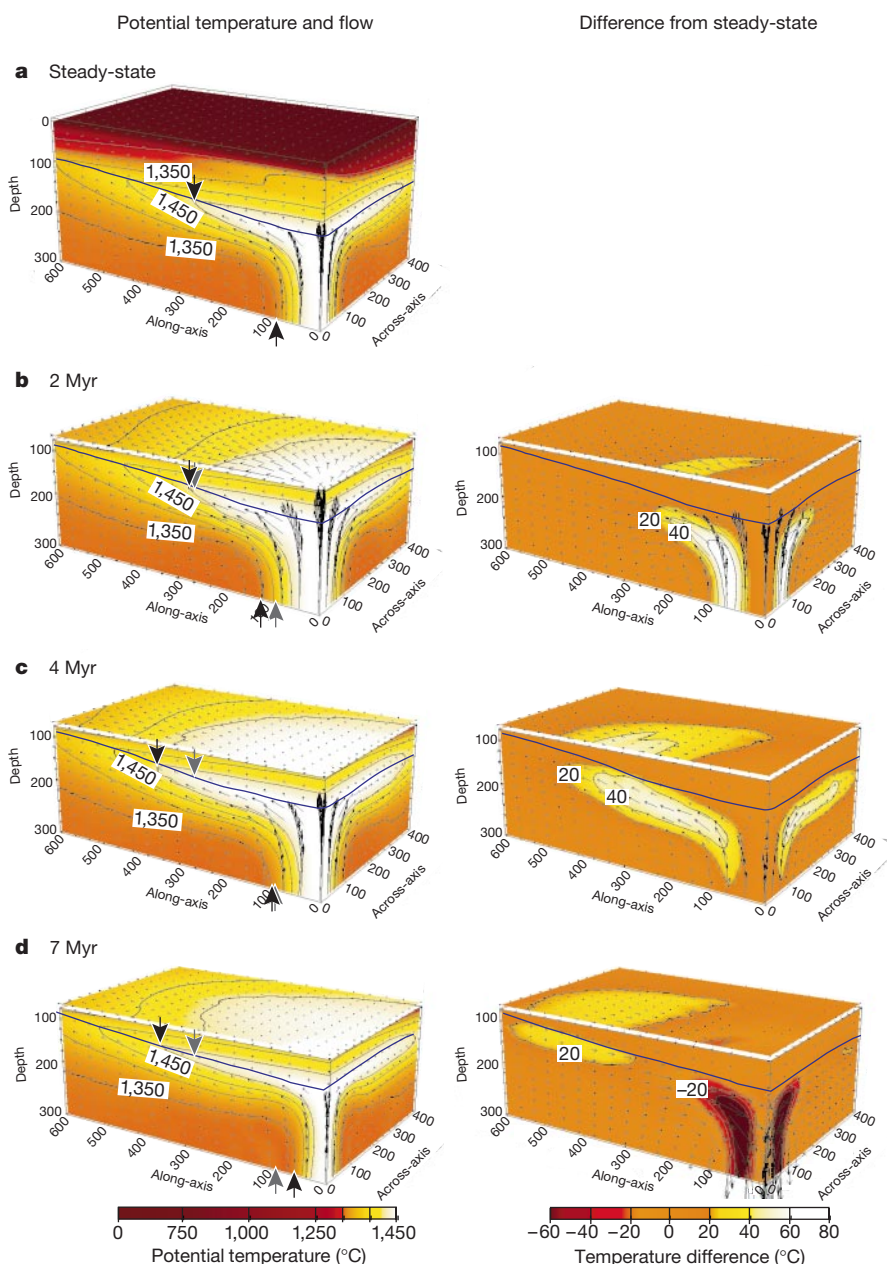


Figure 2 Potential temperatures and flow fields (left) and differences from the steady-state model (right). Potential temperatures (contour interval 100°C) and flow (arrow size proportional to flow rate) are shown on the left, and differences in temperature (contour interval 20°C) and flow from the steady-state plume model (arrows point down if ascent is slower than in steady-state model) are shown on the right. The left vertical panels are along the ridge axis; the right vertical panels are perpendicular to the ridge axis. The box sides are reflecting boundaries and the tops and bottoms (except in the plume) have uniform potential temperatures of 0 and $1,350^{\circ}\text{C}$, respectively. Plate spreading is simulated by imposing a constant horizontal velocity (the half-spreading rate is 9.5 km Myr^{-1} ; ref. 28) at the surface of the box, and the plume is generated by imposing a cylindrical temperature anomaly (decreases as a gaussian function from a maximum of 200°C to 5°C at specified plume radius) in the lower portion of the box. The ambient (reference) viscosity at a depth of 250 km is $5 \times 10^{19}\text{ Pa s}$; the minimum viscosity in the

plume is $5.8 \times 10^{18}\text{ Pa s}$; and the cut-off maximum viscosity is $2.5 \times 10^{22}\text{ Pa s}$. Dehydration viscosity increase (boundary marked by bold blue curve) is incorporated by a pre-exponential multiplying factor¹¹ that increases from 1 to 100 over mantle depletions of 0 to 5%. Melting calculations are based on ref. 29. The boxes have across-axis, along-axis and vertical dimensions $1,152 \times 1,152 \times 500\text{ km}$; number of grids are $162 \times 128 \times 100$. **a**, Steady-state plume beneath ridge axis. Upper bold arrow marks along-axis extent of $1,500^{\circ}\text{C}$ isotherm; lower bold arrow marks boundary of plume stem. **b**, 2 Myr after the initiation of plume flux perturbation. Horizontal cross-section is at a depth of 145 km . Bold arrows show new extent of $1,500^{\circ}\text{C}$ isotherm and plume stem boundary; grey arrows show corresponding positions in the steady-state model (separation between bold and grey arrows indicate changes in temperature). **c**, Time is 4 Myr and horizontal cross-section is at a depth of 135 km . **d**, Time is 7 Myr and horizontal cross-section is at a depth of 130 km .

perpendicularly towards the ridge axis to form crust is illustrated in Fig. 3a. Superimposed on a long-wavelength thinning of the crust away from the plume are shorter-wavelength lineations that angle toward the ridge axis in the form of nested Vs. The maximum variation in crustal thickness occurs at an along-axis distance from the plume centre of about 200 km and decreases to nearly zero at a distance of approximately 600 km. Both the distance of maximum variation and the length of the V-shaped ridges remain nearly constant among models in which plume flux is perturbed by different amounts (the flux range examined here is 46% to 171% of the medians).

The quantity that is most sensitive to changes in flux perturbation is the magnitude of crustal thickness variation. The crustal thickness pattern in Fig. 3a is the predicted by a model in which the plume flux was varied by $(6.1 \pm 2.6) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$, or a total variation of 83% of the indicated median value. The maximum crustal thickness variation in this calculation is 4.9 km, a value slightly greater than, but comparable to, the range of 2–4 km inferred in ref. 4. A predicted linear increase in maximum crustal thickness variation

with normalized flux variation (Fig. 4) suggests that the crustal thickness variation estimated for the North Atlantic requires substantial flux perturbations of the Iceland plume (34–68% of the median flux). The model also predicts an approximately 3-km variation in crustal thickness over Iceland, and indeed there is stratigraphic evidence that crustal productivity on Iceland has varied at periods comparable to those modelled here^{22,23}.

Figure 3b shows the predicted free-air gravity anomaly that results if the crust is compensated by Airy isostasy (see legend for details). The sensitivity of free-air gravity anomalies to the shorter-wavelength variations in crustal thickness yields prominent V-shapes. The predicted gravity lineations have maximum amplitudes of 35 mGal and lengths of about 600 km, both of which are comparable to observations in the North Atlantic⁴.

It is interesting that the crustal and gravity lineations appear nearly straight. Assuming that the crustal Vs in the North Atlantic formed on the ridge axis, the tangent of the angle between a lineation and the ridge axis is the ratio of the ridge half-spreading rate and the along-axis flow rate of a melt perturbation¹. Previous workers have argued that the apparent straightness of the gravity features indicates a constant along-axis flow rate, which would not occur by radial expansion of plume material^{1,4}. Indeed, as suggested by Vogt¹, the models presented here predict the rate of along-axis flow to decrease in proportion to the inverse of radial distance from the plume centre. Consequently, the predicted V-shaped ridges curve. But they do not curve much. The calculation that produced Fig. 3 has an along-axis plume flow rate that is a maximum of 168 km Myr^{-1} at an along-axis distance of 100 km from the plume centre and decreases to 28 km Myr^{-1} at a distance of 600 km. These flow rates predict the angle between a crustal V and the ridge axis to increase from 3° to 19° . This small change in angle is evident in curves computed from the maximum along-axis flow rates, which closely follow the peaks of the crustal Vs (Fig. 3). In the North Atlantic, the gravity lineations are even longer (Fig. 1), estimated flow rates ($\sim 200 \text{ km Myr}^{-1}$; ref. 1) are greater, and thus bends in crustal Vs are likely to be even smaller than predicted by the above model. Indeed, small bends have been identified in sea-floor basement topography¹.

Finally, I discuss the importance of the rheological boundary caused by mantle dehydration. First, the rheological boundary is a requirement for the formation of elongate V-shaped crustal ridges. Numerical calculations that omit dehydration predict dramatic

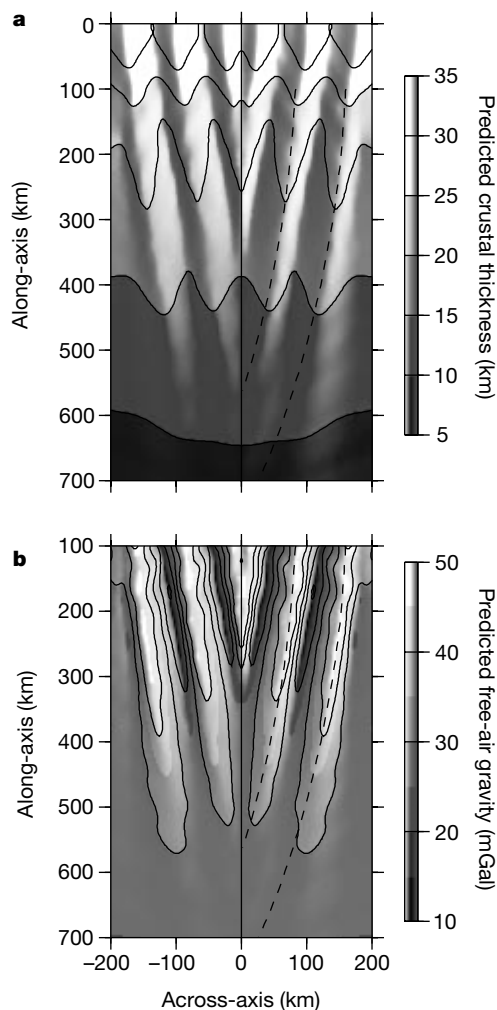


Figure 3 Maps of (a) predicted crustal thickness (contour interval 5 km) and (b) free-air gravity anomaly (contour interval 10 mGal) predicted from crustal structure. Illumination is from the right. Topography of crust–water and crust–mantle interfaces computed by assuming Airy isostatic compensation of crust. The gravitational attraction of the two interfaces was computed³⁰ and then summed. I assumed an upper crustal density of $2,700 \text{ kg m}^{-3}$, a lower crustal density of $3,000 \text{ kg m}^{-3}$ (ref. 4), and a flat interface separating the two. Assumed water and mantle densities are $1,000$ and $3,300 \text{ kg m}^{-3}$, respectively. Solid lines mark ridge axis; dashed curves show trajectory of V-shaped ridges predicted from maximum along-axis flow rates.

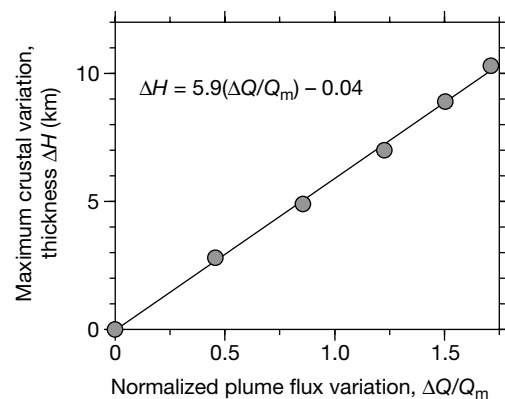


Figure 4 Maximum crustal thickness variations (dots) predicted by models of different variation in plume flux. Flux perturbation is given as difference between maximum and minimum flux ΔQ , normalized by median value Q_m . In order of increasing crustal thickness variation, median fluxes and variations (given as $\pm \Delta Q/2$) are $(5.5 \pm 0) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$; $(5.7 \pm 1.3) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$; $(6.1 \pm 2.6) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$; $(6.2 \pm 3.8) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$; $(6.5 \pm 4.9) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$; $(6.8 \pm 5.8) \times 10^6 \text{ km}^3 \text{ Myr}^{-1}$. Line is the indicated best-fitting function.

variations in upwelling rate above the plume stem, which lead to unrealistically large crustal thickness variations (>200 km). Most of the variation is confined to the plume stem and prominent V-shaped ridges fail to form. Secondly, the depth of the dehydration boundary does not increase significantly away from the ridge axis (Fig. 2a). This prevents the formation of a rheological groove from the off-axis thickening of a thermally controlled lithosphere, which could otherwise channel plume material along the ridge axis for very-low-viscosity plumes²⁴. Flow away from the plume stem is therefore nearly radial. This model thus demonstrates how a radially expanding and pulsing plume can generate Reykjanes V-shaped ridges as well as similar sea-floor age-transgressive ridges south of the Azores hotspot^{25,26}. The possibility of a high-viscosity dehydration layer at other plume-ridge systems has broad implications for the fundamental dynamics of plume-ridge interaction as well as asthenosphere-lithosphere interaction in general. □

Received 13 October 2000; accepted 30 March 2001.

- Vogt, P. R. Asthenosphere motion recorded by the ocean floor south of Iceland. *Earth Planet. Sci. Lett.* **13**, 153–160 (1971).
- Schilling, J.-G. Upper mantle heterogeneities and dynamics. *Nature* **314**, 62–67 (1985).
- Vogt, P. R. Plumes, subaxial pipe flow, and topography along the mid-ocean ridge. *Earth Planet. Sci. Lett.* **29**, 309–325 (1976).
- White, R. S., Brown, J. W. & Smallwood, J. R. The temperature of the Iceland plume and origin of outward propagating V-shaped ridges. *J. Geol. Soc. Lond.* **152**, 1039–1045 (1995).
- Hanan, B. B. & Schilling, J.-G. The dynamic evolution of the Iceland mantle plume: the lead isotope perspective. *Earth Planet. Sci. Lett.* **151**, 43–60 (1997).
- Schilling, J.-G. & Noe-Nygaard, A. Faeroe-Iceland plume: Rare-earth evidence. *Earth Planet. Sci. Lett.* **24**, 1–14 (1974).
- Schilling, J. G., Meyer, P. S. & Kingsley, R. H. Evolution of the Iceland hotspot. *Nature* **296**, 313–320 (1982).
- Hardarson, B. S., Fitton, J. G., Ellam, R. M. & Pringle, M. S. Rift relocation—a geochemical and geochronological investigation of a paleo-rift in northwest Iceland. *Earth Planet. Sci. Lett.* **153**, 181–196 (1997).
- Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: Implications for rheology, melt extraction, and the evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
- Ito, G., Lin, J. & Gable, C. W. Dynamics of mantle flow and melting at a ridge-centered hotspot: Iceland and the Mid-Atlantic Ridge. *Earth Planet. Sci. Lett.* **144**, 53–74 (1996).
- Ito, G., Shen, Y., Hirth, G. & Wolfe, C. Mantle flow, melting, and dehydration of the Iceland mantle plume. *Earth Planet. Sci. Lett.* **165**, 81–96 (1999).
- Gable, C. W. *Numerical Models of Plate Tectonics and Mantle Convection in Three Dimensions*. Thesis, Harvard Univ. (1989).
- Feigener, M. A. & Richards, M. A. The fluid dynamics of plume-ridge and plume-plate interactions: an experimental investigation. *Earth Planet. Sci. Lett.* **129**, 171–182 (1995).
- Ribe, N., Christensen, U. R. & Theissing, J. The dynamics of plume-ridge interaction. 1. Ridge-centered plumes. *Earth Planet. Sci. Lett.* **134**, 155–168 (1995).
- Allen, R. M. *et al.* The thin hot plume beneath Iceland. *Geophys. J. Int.* **137**, 51–63 (1999).
- Wolfe, C., Bjarnason, I. T., VanDecar, J. C. & Solomon, S. C. Seismic structure of the Iceland mantle plume. *Nature* **385**, 245–247 (1997).
- Menke, W., Brandsdóttir, B., Einarsson, P. & Bjarnason, I. Th. Reinterpretation of the RRISP-77 Iceland shear-wave profiles. *Geophys. J. Int.* **126**, 166–172 (1996).
- Staples, R. K. *et al.* Faeroe-Iceland ridge experiment. 1. Crustal structure of northeastern Iceland. *J. Geophys. Res.* **102**, 7849–7866 (1997).
- Vogt, P. R. in *Structure and development of the Greenland-Scotland Ridge, New Methods and Concepts* (eds Bott, M. H. P., Saxov, S., Talwani, M. & Thiede, J.) 191–213 (Plenum, New York, 1983).
- Scott, D. R., Stevenson, D. J. & Whitehead, J. A. Jr Observations of solitary waves in a viscously deformable pipe. *Nature* **319**, 759–761 (1986).
- Olson, P. & Christensen, U. Solitary wave propagation in a fluid conduit within a viscous matrix. *J. Geophys. Res.* **91**, 6367–6374 (1986).
- McDougall, I., Kristjánsson, L. & Saemundsson, K. Magnetostratigraphy and geochronology of northwest Iceland. *J. Geophys. Res.* **89**, 7029–7060 (1984).
- Watkins, N. D. & Walker, G. P. L. Magnetostratigraphy of eastern Iceland. *Am. J. Sci.* **277**, 513–584 (1977).
- Albers, M. Channeling of plume flow beneath mid-ocean ridges. *Earth Planet. Sci. Lett.* **187**, 207–220 (2001).
- Cannat, M. *et al.* Mid-Atlantic Ridge–Azores hotspot interactions: along-axis migration of a hotspot-derived event of enhanced magmatism 10 to 4 Ma ago. *Earth Planet. Sci. Lett.* **173**, 257–269 (1999).
- Vogt, P. R. Global magmatic episodes: new evidence and implications for the steady-state mid-oceanic ridge. *Geology* **7**, 93–98 (1979).
- Sandwell, D. & Smith, W. H. F. Marine gravity from Geosat and ERS-1 altimetry. *J. Geophys. Res.* **102**, 10039–10054 (1997).
- DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophys. Res. Lett.* **21**, 2191–2194 (1994).
- McKenzie, D. & Bickle, M. J. The volume and composition of melt generated by extension of the lithosphere. *J. Petrol.* **29**, 625–679 (1988).
- Parker, R. L. The rapid calculation of potential anomalies. *Geophys. J. R. Astron. Soc.* **31**, 447–455 (1973).
- Wessel, P. & Smith, W. H. F. New version of the Generic Mapping Tools released. *Eos* **76**, 329 (1995).

Acknowledgements

Discussions with D. Lizarralde motivated much of the early work of this study. I also appreciate valuable comments by C. Leshner, J.-G. Schilling and D. Sparks. Figures 1, 3 and 4 were produced using Generic Mapping Tools³¹. Funding was provided by NSF-OCE.

Correspondence and requests for materials should be addressed to the author (e-mail: gito@geology.ucdavis.edu).

Earliest evidence for efficient oral processing in a terrestrial herbivore

Natalia Rybczynski* & Robert R. Reisz†

* Duke University, Department of Biological Anthropology and Anatomy, Box 90383, Durham, North Carolina 27708-0383, USA

† University of Toronto in Mississauga, Department of Zoology, 3359 Mississauga Road, Mississauga, Ontario L5L 1C6, Canada

Herbivores can increase their digestion rate by mechanically reducing particle size through oral trituration¹. Groups of terrestrial vertebrates with the greatest capacity to reduce tough plant foods orally are also the most abundant and diverse, as exemplified by ornithomimid dinosaurs during the Mesozoic and extant artiodactyl and perissodactyl mammals². Thus, the effective oral processing of high-fibre plant material seems to represent an evolutionary innovation of both functional and macroevolutionary significance. However, evidence for oral processing is poorly documented in the fossil record, especially during the initial stages of terrestrial vertebrate diversification^{3,4}. Here we report on the basal anomodont *Suminia getmanovi*, the only known Palaeozoic vertebrate in which unequivocal specializations in its cranium and teeth for high-fibre herbivory are well preserved. We propose that the capacity to comminute tough plant foods was critical to the diversification of anomodonts, the most diverse, widely dispersed and abundant group of Palaeozoic terrestrial vertebrates, and to the onset of modern terrestrial ecosystems.

One of the most significant evolutionary events in the history of life on land is the development of the modern pyramidal trophic structure in which terrestrial plants support a large number of herbivores, which in turn supports a relatively small number of predators⁴. Among vertebrates, this trophic organization was not fully developed until the Upper Permian (260 Myr ago) of the

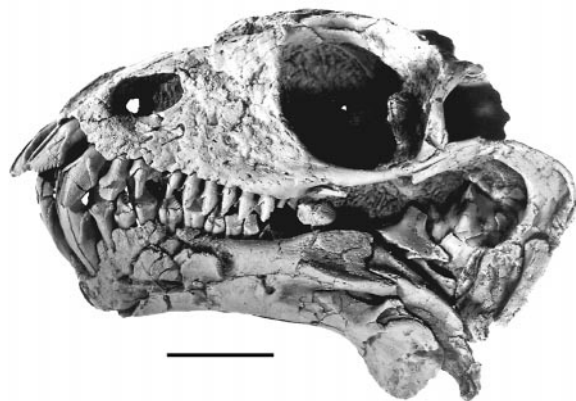


Figure 1 Skull of *Suminia getmanovi*, PIN 2212/62. Scale, 1 cm.

Palaeozoic when the primary consumer base was largely formed by the enormously abundant and diverse anomodonts⁵⁻⁷.

The majority of anomodonts belong to the Dicynodontia, a diverse group characterized by striking cranial specializations including a short snout, an expanded temporal region for adductor musculature, an anteroposteriorly sliding jaw joint (permitting propalinal, or fore-aft, movement, and a beak. This suite of traits has traditionally been thought to have facilitated the processing of plant matter^{8,9}, but they do not provide good evidence of herbivory. For example, birds and turtles, the living beaked vertebrates, both possess herbivorous and non-herbivorous forms, and among turtles both herbivorous and non-herbivorous forms show propalinity². Thus, the presence of beaks and propalinity cannot be directly correlated with herbivory and extensive oral processing. Moreover, it is difficult to infer the relative capacity of dicynodonts to process food orally because their beaks are not preserved in the fossil record and there is no dental occlusion in the few dicynodont taxa that have retained teeth.

*Suminia getmanovi*¹⁰ is a Late Permian (Upper Tatarian) basal anomodont from Kotelnich, in the Vjatka region of Central Russia. Represented by several well preserved skulls (Fig. 1) and four partial skeletons, *Suminia* was a small animal with large eyes, a short, high skull, and long limbs. Its skull is similar to that of dicynodonts, and also possessed an expanded temporal region and sliding jaw joint¹¹.

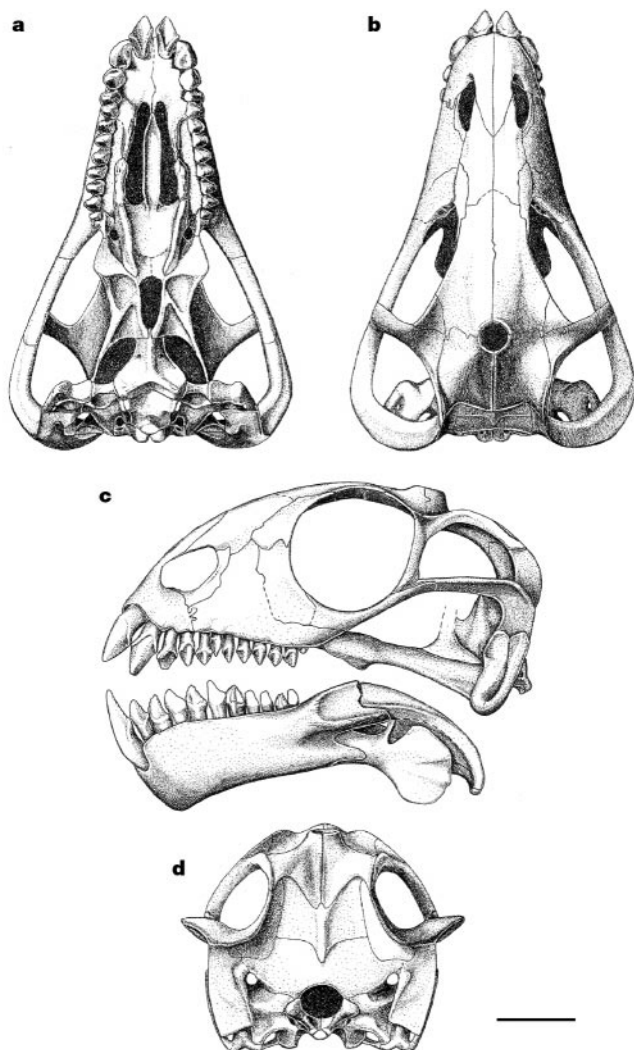


Figure 2 Skull restoration of *Suminia getmanovi*. Based largely on PIN 2212/62, in ventral (a), dorsal (b), lateral (c) and occipital (d) views. Scale, 1 cm.

Instead of a beak, it possessed a complete marginal dentition of 'leaf-shaped' teeth, which are labiolingually compressed with a widened base, and coarse serrations along anterior and posterior edges that are oriented at 45° or less from the margin of the tooth¹². Such teeth have evolved convergently in at least five independent lineages of terrestrial herbivores: iguanid lizards, ornithischian and prosauropod dinosaurs, pareiasaurs and caseid synapsids^{4,6,12}. In *Suminia*, the teeth are closely packed and oriented in echelon so that the serrated edges of the teeth are angled obliquely relative to the long axis of the tooth row. A similar condition is seen in iguanas, ornithischian dinosaurs and prosauropod dinosaurs¹².

The overall pattern of tooth shape and wear in *Suminia* is entirely consistent with herbivory. The enlarged anterior incisiform teeth were probably used to detach portions of plants for ingestion, in a manner similar to the herbivorous lizard *Uromastix*¹³. The presence of anterior teeth specialized for cropping is indicative of an herbivorous diet because herbivores must be able to detach portions

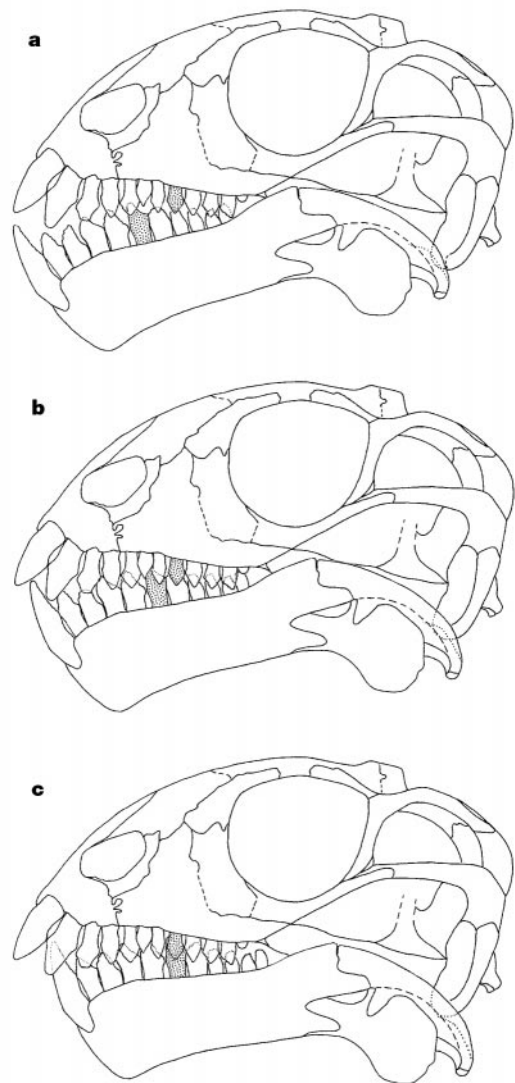


Figure 3 Reconstructed power stroke in *Suminia*. Stippled teeth in the upper jaw and dentary bones are U8 and L6, respectively. The power stroke can be divided into three major stages. **a**, In the earliest phase of occlusion the lower teeth move along the lingual surface of the upper teeth in a motion that combines elevation and retraction. **b**, In the second stage the direction of movement of the lower teeth becomes increasingly longitudinal. **c**, In the last stage of the power stroke the jaw motion is nearly parallel with the tooth row.

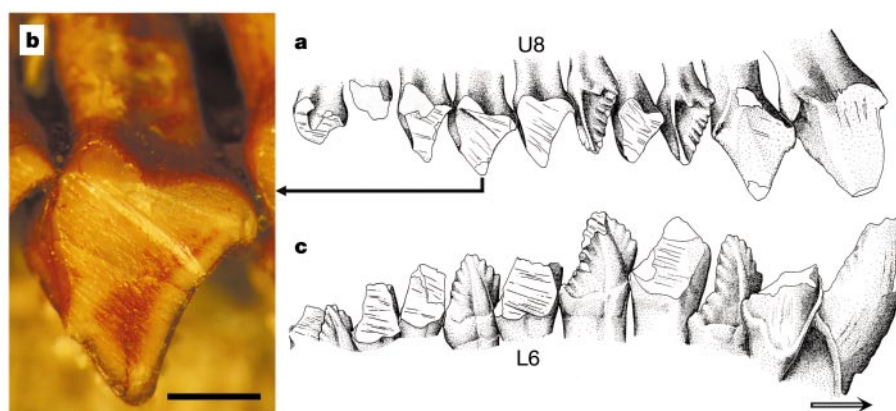


Figure 4 Lingual surface of the right upper tooth row and buccal surface of the right lower tooth row of *Suminia* (PIN 2212/62) and photomicrograph image of U8 tooth. **a**, The upper tooth row and **b**, **c**, U8 are shown in mirror-image so that they may be compared directly to those of the lower tooth row. The wear facet of U8 is characterized by a large ventral region of steeply-angled ($50 \pm 5^\circ$) striations and a smaller region of shallower-angled features ($20 \pm 5^\circ$). These two regions are separated by a step, with the ventral region

being recessed. The dorsal striations were created by L6 whereas the ventral striations were created by the more laterally positioned L8. The step between the two regions is the result of L8 being more laterally positioned than L6. L7 is newly erupted, non-occluding lower tooth that accounts for the step seen in U8. The arrow with double-lined shaft indicates the anterior direction on the dentition. Scale, 1 mm.

of plants before ingestion. *Suminia*'s posterior teeth processed ingested food materials, and the upper and lower posterior tooth rows have wear facets on the labial and lingual aspects of the teeth, respectively. These wear facets are distinct and evidently derived from precise, tooth-to-tooth occlusion, rather than tooth-to-food wear¹⁴.

The jaw action that processes food within the oral cavity is the power stroke (Fig. 3). In *Suminia*, the high angle of the occlusal plane (about $7.0 \pm 5^\circ$ from the plane of the jaws) and the presence of roughly parallel striations over the entire surface of the wear facets (Fig. 4) indicate that food particles were reduced with a precise, shearing power stroke (as opposed to crushing). Such an occluding, leaf-shaped dentition with a shearing power stroke is not seen again until it appears convergently in the Mesozoic with the appearance of ornithischian dinosaurs.

The superb preservation of the skull and dentition of *Suminia* (Figs 1, 2) made it possible to infer the pattern of tooth occlusion during the power stroke, the first such reconstruction for a Palaeozoic tetrapod. The shape of the jaw joint and the steeply angled wear facets indicate that jaw movement during the power stroke was constrained to the sagittal plane. The direction of the power stroke was inferred from a combination of evidence including: (1) the posteriorly oriented external adductor musculature; (2) the wear profiles across the enamel–denture interfaces¹⁵; (3) the concave curvature of the anterior blade edges of the upper teeth and posterior blade edges of the lower teeth and the occurrence of thickened enamel on the anterior edge of the upper tooth blades^{16,17}. All evidence is consistent with a retractive power stroke.

The pattern of tooth occlusion during the power stroke was reconstructed from microwear striations on the tooth wear facets (Fig. 4). The power stroke would have involved roughly the posterior two-thirds of the tooth row, the jaw would have retracted for about 20% of the total tooth row length, and each of the occluding lower teeth would have contacted two or three upper teeth (Fig. 3). Consequently, during a single power stroke the blade edge of each lower tooth would have occluded with several upper teeth (and vice versa). This contrasts with an orthal power stroke in which each blade edge functions once per power stroke. Thus, the retractive power stroke in *Suminia* permitted an enhanced capacity to reduce mechanically tough plant material.

In summary, the dentition of *Suminia* is clearly more specialized for high-fibre herbivory than any other known Palaeozoic tetrapod,

and the combined cranial and dental anatomy provides conclusive evidence that extensive oral processing took place in this anomodont. This finding is significant not only because it is the oldest known such example in the evolutionary history of terrestrial vertebrates, but also because it indicates an association between the evolution of the anomodont/dicynodont-type cranial architecture and the extensive oral processing of high-fibre plant materials. Specifically, *Suminia* is a basal member of Anomodontia¹¹, and this suggests that the capacity to process orally tough plant material may have been a basal 'innovation' within anomodonts. In conclusion, effective oral processing of high-fibre plant materials appears to have been an important factor in the evolution and diversification of the anomodonts, the largest herbivorous group that was associated with the onset of the modern terrestrial ecosystem. □

Received 10 January; accepted 21 March 2001.

1. Bjorndal, K. A., Bolton, A. B. & Moore, J. E. Digestive fermentation in herbivores: Effect of food particle size. *Phys. Zool.* **63**, 710–721 (1990).
2. Norman, D. B. & Weishampel, D. B. Ornithomorph feeding mechanisms: Their bearing on the evolution of herbivory. *Am. Nat.* **126**, 151–164 (1985).
3. King, G. M. *Reptiles and Herbivory* (Chapman and Hall, London, 1996).
4. Reisz, R. R. & Sues, H.-D. in *Evolution of Herbivory in Terrestrial Vertebrates* (ed. Sues, H.-D.) 9–41 (Cambridge Univ. Press, Cambridge, 2000).
5. Di Michelle, W. A. & Hook, R. W. in *Terrestrial Ecosystems Through Time* (eds Behrensmeyer, A. K. et al.) 205–325 (Univ. Chicago Press, Chicago, 1992).
6. Sues, H.-D. & Reisz, R. R. Origins and early evolution of herbivory in tetrapods. *Trends Ecol. Evol.* **18**, 141–145 (1998).
7. Hotton, N., Olson, E. C. & Beerbower, R. in *Amniote Origins: Completing the Transition to Land* (eds Sumida, S. S. & Martin, K. L. M.) 207–264 (Academic, San Diego, 1997).
8. Hotton, N. in *The Ecology and Biology of Mammal-like Reptiles* (eds Hotton, N., Maclean, P. D., Roth, J. J. & Roth, E. C.) 71–82 (Smithsonian Institution Press, Washington, 1986).
9. King, G. M. The early anomodont *Venjukovia* and the evolution of the anomodont skull. *J. Zool. (Lond.)* **232**, 651–673 (1994).
10. Ivachnenko, M. F. A new Late Permian dromosaurian (Anomodontia) from Eastern Europe. *Paleontol. J.* **28**, 96–103 (1994).
11. Rychczynski, N. Cranial anatomy and phylogenetic position of the basal anomodont (Therapsida), *Suminia getmanovi*. *Zool. J. Linn. Soc.* **130**, 329–373 (2000).
12. Galton, P. M. in *The Beginning of the Age of Dinosaurs* (ed. Padian, K.) 203–221 (Cambridge Univ. Press, Cambridge, 1986).
13. Throckmorton, G. Oral food processing in two herbivorous lizards, *Iguana iguana* (Iguanidae) and *Uromastix aegyptius* (Agamidae). *J. Morphol.* **148**, 363–390 (1976).
14. Popowicz, T. E. & Fortelius, M. On the cutting edge: Tooth blade sharpness in herbivorous and faunivorous mammals. *Ann. Zool. Fennici* **34**, 73–88 (1997).
15. Costa, R. L. & Greaves, W. S. Experimentally produced tooth wear facets and the direction of jaw motion. *J. Paleontol.* **55**, 635 (1981).
16. Rensberger, J. M. An occlusion model for mastication and dental wear in herbivorous mammals. *J. Paleontol.* **47**, 515–528 (1973).
17. Rensberger, J. M. Function in the cheek tooth evolution of some hypsodont geomyid rodents. *J. Paleontol.* **49**, 10–22 (1975).

Acknowledgements

Specimens of *Suminia* were prepared and loaned to us by M. F. Ivachnenko and A. Khlupin. We thank H.-D. Sues, K. Smith, W. Hylander and J. Rensberger for comments on earlier versions, and D. Scott for assistance with the figures. Support for this research was provided by the National Geographic Society and Natural Sciences and Engineering Research Council of Canada.

Correspondence and requests for materials should be addressed to R.R.R. (e-mail: rreis@credit.erin.utoronto.ca).

Effects of macrophyte species richness on wetland ecosystem functioning and services

Katharina A. M. Engelhardt* & Mark E. Ritchie

Department of Fisheries and Wildlife and The Ecology Center, 5210 Old Main Hill, Utah State University, Logan, Utah 84322-5210, USA

Wetlands provide many important ecosystem services to human society^{1–5}, which may depend on how plant diversity influences biomass production and nutrient retention^{4,6–8}. Vascular aquatic plant diversity may not necessarily enhance wetland ecosystem functioning, however, because competition among these plant species can be strong, often resulting in the local dominance of a single species^{4,9}. Here we have manipulated the species richness of rooted, submerged aquatic plant (macrophyte) communities in experimental wetland mesocosms. We found higher algal and total plant (algal plus macrophyte) biomass, as well as lower loss of total phosphorus, in mesocosms with a greater richness of macrophyte species. Greater plant biomass resulted from a sampling effect; that is, the increased chance in species mixtures that algal production would be facilitated by the presence of a less competitive species—in this case, crisped pondweed. Lower losses of total phosphorus resulted from the greater chance in species mixtures of a high algal biomass and the presence of sago pondweed, which physically filter particulate phosphorus from the water^{2,10,11}. These indirect and direct effects of macrophyte species richness on algal production, total plant biomass and phosphorus loss suggest that management practices that maintain macrophyte diversity may enhance the functioning and associated services of wetland ecosystems.

A critical question in environmental biology is whether macrophyte diversity in wetlands determines the effectiveness of the well-known services of wetlands to society, such as the sustainable production of food, recreational opportunities, and water purification by retention of pollutants and sediments. These services probably depend on how well wetlands perform certain ecosystem functions, such as nutrient retention^{1,2,12} and primary production^{1,13,14}. Work in grasslands has suggested that greater plant species richness leads to more efficient uptake of nutrients and greater productivity^{15–18}, however, local environments in wetlands are typically dominated by a single vascular plant species^{4,9}. Thus, vascular plant diversity in wetlands may not affect ecosystem functioning positively, or even by the same mechanisms operating in grassland systems, and therefore biodiversity may not positively affect ecosystem functioning ubiquitously. For these reasons, we investigated whether the diversity of submerged, rooted freshwater

aquatic vascular plants can affect wetland biomass production and phosphorus retention—two ecosystem processes closely related to wetland ecosystem services^{1,2}.

We manipulated species richness of four submerged aquatic macrophyte species, sago pondweed (*Potamogeton pectinatus*), long-leaved pondweed (*Potamogeton nodosus*), crisped pondweed (*Potamogeton crispus*) and horned pondweed (*Zannichellia palustris*), in experimental mesocosms. The species are functionally and morphologically different, for example in their use of space and resources in soil, water and air.

Aboveground biomass of macrophytes ('shoot biomass') was measured to understand how macrophyte biomass is correlated with macrophyte species richness. Periphyton, which were present predominantly as green filamentous algae and hereafter are referred to as 'algae', were also measured because the macrophyte species differed in how well they supported algae, and because algae are an important structural and functional component of wetlands^{11,19}. Phytoplankton were not measured owing to their relatively low biomass compared with the biomass produced by filamentous algae.

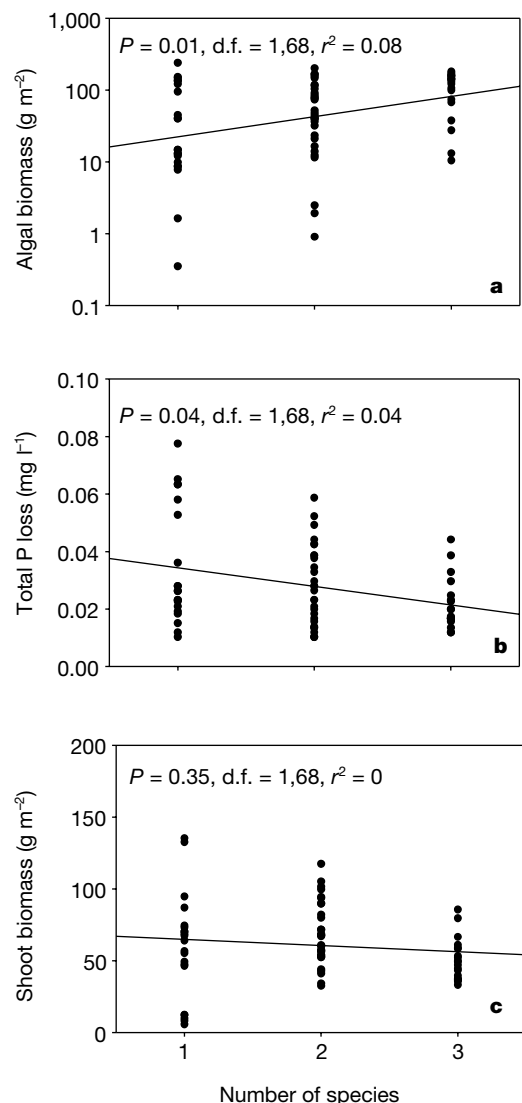


Figure 1 The effect of species richness on algal biomass (a), nutrient retention (b) and above-ground macrophyte biomass (c) (mean $\pm$ s.e.). Solid line is regression of biomass or total P versus species richness. Algal biomass is periphyton biomass that is mostly composed of green filamentous algae. Nutrient retention was inferred from measured total phosphorus (P) loss from the outflow of each mesocosm. Shoot biomass is aboveground biomass of submerged aquatic macrophytes.

* Present address: University of Maryland, Center for Environmental Science, Appalachian Laboratory, 301 Braddock Road, Frostburg, Maryland 21532-2307, USA.

Table 1 Algal biomass and total phosphorus loss in the presence and absence of the four species of pondweed

	Sago		Long-leaved		Crisped		Horned	
	Presence	Absence	Presence	Absence	Presence	Absence	Presence	Absence
Algal biomass (g m^{-2})	62.0* $\pm$ 10.2	95.2 $\pm$ 10.5	85.1 $\pm$ 9.7	72.1 $\pm$ 11.6	125.3*** $\pm$ 7.9	31.9 $\pm$ 6.2	70.7 $\pm$ 10.5	86.6 $\pm$ 10.8
Total P loss (mg l^{-1})	0.017 $\pm$ 0.001	0.022 $\pm$ 0.003	0.018 $\pm$ 0.001	0.021 $\pm$ 0.003	0.016** $\pm$ 0.001	0.023 $\pm$ 0.003	0.023* $\pm$ 0.003	0.016 $\pm$ 0.001

Values are means $\pm$ s.e.m. Stars in the 'presence' columns indicate statistically significant differences among means when a particular species is present versus when it is absent from the communities: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Nutrient loss was measured as loss of total phosphorus (P) in the outflow of the mesocosms. Thus, our study measured nutrient loss directly, rather than indirectly by measuring nutrient concentrations in the soil below a perceived rooting zone¹⁶. Total P is often used to measure a wetland's nutrient status and ability to retain nutrients^{2,20}, and includes primarily P bound-up in particulate matter (80% of total P in our mesocosm systems) rather than dissolved mineral P directly available to plants.

Algal biomass (Fig. 1a) increased and total P loss (Fig. 1b) decreased with increasing macrophyte species richness. Because macrophyte shoot biomass did not change with macrophyte species richness (Fig. 1c) but algal biomass did, total plant biomass (macrophyte plus algae) also increased with macrophyte species richness (linear regression, $P = 0.06$, $R^2 = 0.04$).

Greater macrophyte species richness resulted in higher algal biomass, and thus total plant biomass, because of an indirect 'sampling effect'²¹. Higher macrophyte species richness increased the chance that crisped pondweed would occur in the community, and thus seemed indirectly to increase the chance that a mesocosm would support higher algal biomass. Crisped pondweed was associated with the highest production of algal biomass (Table 1), presumably by providing attachment space for algae and nutrients for algal growth through leaching of leaf tissue^{22–24}.

Notably, this indirect sampling effect overcame an 'inverse sampling effect'^{25,26} of the greater chance in species mixtures that a competitively dominant but less productive macrophyte species—sago pondweed—could be present. The competitive dominance of sago pondweed (Fig. 2a), as judged from its higher than expected

biomass per genet in species mixtures, apparently caused macrophyte tricultures to converge towards the biomass of sago pondweed in monoculture (Fig. 1c). Therefore, even though crisped pondweed yielded lower than expected biomass per genet in species mixtures with sago pondweed present (Fig. 2c), its facilitation of algae was sufficient to produce a sampling effect towards higher algal biomass (Fig. 1a) and higher plant (algal plus macrophyte) biomass.

Phosphorus loss in the outflow correlated negatively with algal biomass ($P < 0.001$, $r^2 = 0.30$) and decreased substantially when crisped and sago pondweed were present (Table 1). Thus, greater species richness decreased total P loss in the mesocosm outflow (Fig. 1b) through the indirect sampling effect by crisped pondweed on algal biomass, but also through a direct sampling effect by sago pondweed. Because total P is largely immobilized P that is bound-up in particulate matter, lower P losses should occur primarily because plants physically filter particulate matter from the water rather than from uptake of available P. Sago pondweed may be important in filtering P because of its highly reticulate structure. Thus, when algal biomass is high owing to facilitation by crisped pondweed, or when sago pondweed is present in a community, physical filtration of particulate phosphorus seems to be enhanced^{2,10,11}.

Higher productivity from greater species diversity sometimes arises from the greater chance in species mixtures that superior competitors with high productivity are present in a community²¹. Such sampling effects are sometimes construed as evidence that individual species, rather than species richness, influence ecosystem functioning^{27,28}. In our case, however, the species with the greatest indirect effects on algal biomass—crisped pondweed—was a weaker competitor than sago pondweed. Crisped pondweed's biomass per genet in species mixtures with sago pondweed was less than its biomass per genet in monoculture (Fig. 2c), whereas the opposite was true for sago pondweed (Fig. 2a). Furthermore, P losses were reduced by both the presence of crisped and sago pondweed, even though the sampling effect by sago pondweed did not increase macrophyte biomass.

Thus, our results emphasize the importance of species diversity in enhancing ecosystem functioning, because the plant species with the biggest effect on algal biomass and P retention—crisped pondweed—was not a superior competitor and would probably be excluded from a community because of interspecific competition. In addition, the sampling effect of sago pondweed on total P combines with the indirect sampling effect of crisped pondweed to produce a diversity effect in conjunction with significant species effects. These probable mechanisms, and the fact that the four macrophyte species were competing strongly (Fig. 2), effectively refute a null hypothesis that greater species richness merely increased the chance of sampling faster-growing, and thus more productive, species in the absence of competition²⁸.

Our results imply that higher vascular plant species richness in wetlands may potentially yield up to 25% more algal biomass, thereby potentially supporting a greater abundance of fish and wildlife, and retaining up to 30% more potentially polluting nutrients, such as P. These results are important because many wetlands are dominated by one or a few vascular plant species. In fact, managed freshwater wetlands near the Great Salt Lake, Utah, are, in the absence of disturbances, predominantly monocultures of

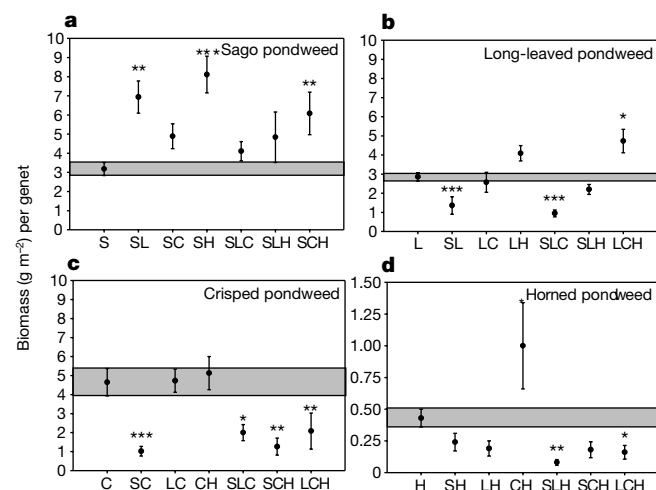


Figure 2 Biomass yield (mean $\pm$ s.e.) per genet (shoot biomass per individual planted) of the four submerged macrophyte species (a–d) in monocultures and in mixed cultures. C, crisped pondweed; H, horned pondweed; L, long-leaved pondweed; S, sago pondweed. Biomass per genet standardizes the number of individuals planted in different mixtures (21 individuals in monocultures, 10 individuals per species in bicultures, and 7 individuals per species in tricultures). A species experienced interspecific competition if its biomass per genet was significantly lower ($P < 0.05$) in the presence of other macrophyte species than in monoculture. * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$, in contrasts after analysis of variance.

sago pondweed. This suggests that sago pondweed may ultimately exclude the other macrophyte species in the field, and therefore may decrease the algal and total plant biomass of a wetland. Our results imply that management practices that maintain the diversity of aquatic macrophytes in wetlands, such as sustaining or restoring a natural disturbance regime²⁹ to prohibit exclusion of less competitive species, may sustain ecosystem functioning and promote the services of those wetlands to humans. □

Methods

Experimental design

The experiment consisted of 70 wading pools (1.5-m diameter, 0.5-m high) at the Aquatic Ecology Research Complex in Millville, Utah, filled to 25 cm with local terrestrial soil. Stream water inflow was 2 l h⁻¹, with a retention time of 2 d. We planted macrophytes as either propagules or shoots in the pools in early May 1999, with 11 individuals per square metre and equal numbers of individuals per species. We planted 14 treatments: 5 replicates each of 4 monocultures, 6 bicultures and 4 tricultures, representing all possible one-, two- and three-species combinations of the four submerged aquatic macrophyte species (sago, long-leaved, crisped and horned pondweed). In mixed communities, individuals of any one species were never surrounded by individuals of its own species. All species are native to the USA, except crisped pondweed, which has been naturalized in the USA for over 150 yr.

Biomass and P measurements

We harvested macrophyte shoot biomass and algal biomass (dried at 60 °C and weighed) at the end of August. Total phosphorus was measured by autoclaving an unfiltered water sample with a persulphate/sulphuric acid oxidant, and analysing the autoclaved sample photometrically using the ascorbic acid method³⁰. Competition was inferred when shoot biomass per individual planted, or genet, was significantly lower in the presence of other macrophyte species than in monoculture.

Received 8 March; accepted 26 March 2001.

- Mitsch, W. J. & Gosselink, J. G. *Wetlands* (Van Nostrand Reinhold, New York, 1993).
- Mitsch, W. J., Cronk, J. K., Wu, X., Naiman, R. W. & Hey, D. L. Phosphorus retention in constructed freshwater riparian marshes. *Ecol. Appl.* **5**, 830–845 (1995).
- Costanza, R. *et al.* The value of the world's ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
- Chambers, R. M., Meyerson, L. A. & Saltonstall, K. Expansion of *Phragmites australis* into tidal wetlands of North America. *Aquat. Bot.* **64**, 261–273 (1999).
- Wilson, M. A. & Carpenter, S. R. Economic valuation of freshwater ecosystem services in the United States: 1971–1997. *Ecol. Appl.* **9**, 772–783 (1999).
- Pollock, M. M., Naiman, R. J. & Hanley, T. A. Plant species richness in riparian wetlands—A test of biodiversity theory. *Ecology* **79**, 94–105 (1998).
- Bedford, B. L., Walbridge, M. R. & Aldous, A. Patterns in nutrient availability and plant diversity of temperate North American wetlands. *Ecology* **80**, 2151–2169 (1999).
- Grace, J. B. The factors controlling species density in herbaceous plant communities: An assessment. *Persp. Plant Ecol. Evol. Syst.* **2**, 1–28 (1999).
- Fischer, J. M., Klug, J. L., Reed, A. T. & Chalmers, A. G. Spatial pattern of localized disturbance along a southeastern salt marsh tidal creek. *Estuaries* **23**, 565–571 (2000).
- Wetzel, R. G. Land–water interfaces: Metabolic and limnological regulators. *Verh. Internat. Verein. Limnol.* **24**, 6–24 (1990).
- Wu, X. & Mitsch, W. J. Spatial and temporal patterns of algae in newly constructed freshwater wetlands. *Wetlands* **18**, 9–20 (1998).
- Jeppesen, E. *et al.* Lake and catchment management in Denmark. *Hydrobiologia* **395/396**, 419–432 (1999).
- Bunn, S. E. & Boon, P. I. What sources of organic carbon drive food webs in billabongs? A study based on stable isotope analysis. *Oecologia* **96**, 85–94 (1993).
- Hargeby, A., Andersson, G., Blindow, I. & Johansson, S. Trophic web structure in a shallow eutrophic lake during a dominance shift from phytoplankton to submerged macrophytes. *Hydrobiologia* **279–280**, 83–90 (1994).
- Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Empirical evidence that declining species diversity may alter the performance of terrestrial ecosystems. *Phil. Trans. R. Soc. Lond. B* **347**, 249–262 (1995).
- Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
- Symstad, A. J., Tilman, D., Willson, J. & Knops, J. M. H. Species loss and ecosystem functioning: effects of species identity and community composition. *Oikos* **81**, 389–397 (1998).
- Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
- Robinson, G. G. C., Gurney, S. E. & Goldsborough, L. G. The primary productivity of benthic and planktonic algae in a prairie wetland under controlled water-level regimes. *Wetlands* **17**, 182–194 (1997).
- Carlson, R. E. A trophic state index for lakes. *Limnol. Oceanogr.* **22**, 361–369 (1977).
- Tilman, D., Lehman, C. L. & Thomson, K. T. Plant diversity and ecosystem productivity: theoretical considerations. *Proc. Natl Acad. Sci. USA* **94**, 1857–1861 (1997).
- Moeller, R. E., Burkholder, J. M. & Wetzel, R. G. Significance of sedimentary phosphorus to a rooted submersed macrophyte (*Najas flexilis* (Willd.) Rostk. & Schmidt) and its algal epiphytes. *Aqu. Bot.* **32**, 261–281 (1988).
- Burkholder, J. M. & Wetzel, R. G. Microbial colonization on natural and artificial macrophytes in a

phosphorus-limited, hardwater lake. *J. Phycol.* **25**, 55–65 (1989).

- Burkholder, J. M. & Wetzel, R. G. Epiphytic alkaline phosphatase on natural and artificial plants in an oligotrophic lake: Re-evaluation of the role of macrophytes as a phosphorus source for epiphytes. *Limnol. Oceanogr.* **35**, 736–747 (1990).
- Loreau, M. Biodiversity and ecosystem functioning: recent theoretical advances. *Oikos* **91**, 3–17 (2000).
- Troumbis, A. Y., Dimitrakopoulos, P. G., Siamantziouras, A.-S. D. & Memtsas, D. Hidden diversity and productivity patterns in mixed Mediterranean grasslands. *Oikos* **90**, 549–559 (2000).
- Wardle, D. A. Is “sampling effect” a problem for experiments investigating biodiversity–ecosystem function relationships? *Oikos* **87**, 403–407 (1999).
- Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
- Ward, J. V. Riverine landscapes: biodiversity patterns, disturbance regimes, and aquatic conservation. *Biol. Cons.* **83**, 269–278 (1998).
- Wetzel, R. G. & Likens, G. E. *Limnological analyses*, Second edition (Springer, New York, 1991).

Acknowledgements

We thank J. Chase, R. Hilderbrand, L. Pitelka, A. Symstad, D. Tilman and R. G. Wetzel for comments; and R. Chi, C. Hendrix, R. Young, E. Toman, M. Amacher and many volunteers and temporary technicians for field and laboratory help. Funding was provided by the US Fish and Wildlife Service, Utah Division of Wildlife Resources, Utah State University Ecology Center and the Society of Wetland Scientists.

Correspondence and requests for materials should be addressed to K.E. (e-mail: engelhardt@al.umces.edu).

Sexual selection and the maintenance of sex

Steven Siller

Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Sex is expensive. A population of females that reproduce asexually should prima facie have twice the growth rate of an otherwise equivalent anisogamous sexual population lacking paternal care, or a population with modes of paternal care that can be co-opted by parthenogenetic females^{1–6}. The two leading theories for the maintenance of sex require either synergistic interactions between deleterious mutations, or antagonistic epistasis between beneficial mutations⁵. Current evidence is equivocal as to whether the required levels of epistasis exist^{6–10}. Here I show that a third factor, differential male mating success (or, more generally, higher variance in male than in female fitness), can drastically reduce mutational load in sexual populations with or without any form of epistasis. Differential mating success has the further advantage of being ubiquitous, and is likely to have preceded or evolved concurrently with anisogamy¹¹.

The idea that differential male mating success or harsher selection on males in general may have something to do with the maintenance of sex is not new; it dates back at least to Trivers¹² (see also ref. 3). Trivers, however, suggested that sex provided an immediate benefit: females choose males who would give them daughters twice as fit as they would obtain if they had mated randomly. If deleterious mutations are the chief source of additive genetic variance in fitness then this idea is untenable as it requires mutation rates 20–100 times higher than those observed in nature. Other studies have focused on males and their role in the reduction of mutational load in sexual populations^{13–16}. In particular, the argument developed in this paper was presented by Manning¹⁶, but the analysis was restricted to the case of a single locus. Nevertheless, probably owing to the lack of convincing quantitative arguments, differential male mating success has been more or less ignored in recent discussions of the maintenance of sex^{3–6}.

Under the well supported assumption that most deleterious mutations are partially recessive, the genetic load (that is, proportionate reduction from maximum possible fitness), L , for large

asexual populations, regardless of the relationship between number of deleterious mutations and fitness, is to a close approximation,

$$L = 1 - e^{-U} \quad (1)$$

where U is the mean number of new deleterious mutations per diploid genome per generation¹⁷.

When the effects on viability across loci of deleterious mutations are assumed to be independent (multiplicative), the same relationship holds for sexual populations with random mating¹⁷. Moreover, the number of deleterious mutations borne by individuals in the population shows an approximately Poisson distribution at equilibrium. This result may be shown rigorously by application of the method of generating functions¹⁸. However, if one accepts the reasonable assumption that the randomizing effect of meiosis will lead to approximate linkage equilibrium between alleles of small effect and hence rare, slightly deleterious mutations having a roughly Poisson distribution, the result may be derived simply by consideration of average gains and losses of mutations per generation.

To see this, assume that the average relative effect on an individual's viability per deleterious mutation is $1 - s$ when homozygous, and $1 - hs$ when heterozygous. If deleterious alleles have a Poissonian distribution in the population with mean m (that is, the probability that an individual bears i mutations is $m^i e^{-m}/i!$) and the rare instances of homozygous pairs of deleterious alleles are ignored, then after selection the probability of bearing i mutations is

$$\frac{(1 - hs)^i m^i e^{-m}/i!}{\sum_{i=0}^{\infty} (1 - hs)^i m^i e^{-m}/i!} = (1 - hs)^i m^i e^{-m(1 - hs)}/i!$$

where the factor $(1 - hs)^i$ represents the effect of selection and the denominator is the appropriate normalizing function. Selection, under the assumptions of initial Poisson distribution of mutations and their independent effect on fitness, thus reduces the mean number of deleterious mutations in the population by a factor $1 - hs$, but leaves their Poisson distribution across individuals unchanged. Thus, if the population bears on average m_i mutations per individual in generation i , it will bear

$$m_{i+1} = (1 - hs)m_i + U$$

in generation $i + 1$, where U is the mutation rate per generation as before. Solving the above equation, the average number of deleterious mutations borne by an individual will be:

$$\bar{m} = \frac{U}{hs} \quad (2)$$

Mean viability is then approximately $(1 - hs)^{U/hs} \approx e^{-U}$ as individual fitness is approximately additive over the range of genotypes in the population. This gives a genetic load identical to the load for asexuals given in equation (1).

Differential male mating success changes the genetic 'viability' load of a sexual population considerably. To simplify the exposition, I consider 'best of n ' mate choice in which a female samples n males randomly from the population and then mates with the individual that she perceives to be the 'best' in some trait t . Assuming Gaussian distribution of m and t , a female that follows such a rule will mate with a male who has, on average, $\alpha_n \rho_{mt}$ standard deviations fewer mutations than the average male in the population, where ρ_{mt} is the correlation between an individual's actual number of mutations and his perceived level of trait or female preferred resource (for example, male territory quality), and α_n is the leading rankit¹⁹ for a sample of n ($\alpha_{1, \dots, 6} \approx 0, 0.56, 0.85, 1.03, 1.16, 1.27$ and so on).

If we assume that under non-random mating mutations still have a nearly Poisson distribution then the following recurrence relation relates the average number of deleterious mutations per zygotic genome in generation $i + 1$ to the average number in generation i :

$$m_{i+1} \approx (1 - hs)m_i - \frac{1}{2} \alpha_n \rho_{mt} \sqrt{m_i} + U \quad (3)$$

where the second term on the right side of the equation represents

the loss due to mate choice. The factor $1/2$ represents the fact that an offspring inherits only half its genes from its father, and $\sqrt{m_i}$ is the standard deviation in the number of mutations derived from the Poisson assumption. Although equation (3) is derived for best of n mate choice it may be interpreted much more widely, as it is the general dynamic equation for any form of differential male mating success where the average father is $\alpha \rho$ genotypic standard deviations above the average male in the breeding population.

Solving equation (3) for equilibrium m we can obtain the solution in two useful forms: the implicit form,

$$\bar{m} \approx \frac{U}{hs + \frac{\alpha_n \rho_{mt}}{2\sqrt{\bar{m}}}} \quad (4)$$

and the explicit form,

$$\bar{m} \approx \left(\sqrt{\left(\frac{\alpha_n \rho_{mt}}{4hs} \right)^2 + \frac{U}{hs}} - \frac{\alpha_n \rho_{mt}}{4hs} \right)^2 \quad (5)$$

where it has been assumed that ρ_{mt} is its equilibrium value. Equation (4) is clearly the equivalent of equation (2), the random mating result, where hs is the contribution of viability to the selection against suboptimal alleles while $\alpha_n \rho_{mt}/2\sqrt{\bar{m}}$ is the contribution to selection against suboptimal alleles of mate choice.

The Poisson distribution assumption and the resulting equation (5) have been checked for a wide range of reasonable parameter values by computer simulations of large finite populations under the assumption of 'bean-bag genetics' (each individual bears a number of deleterious mutations of independent effect, each of which is independently inherited by offspring with 50% chance—program available on request). The results agreed with those obtained by the mean effect method used in this paper to within a few per cent.

Substituting the reasonable²⁰ values $U = 1$ and $hs = 0.01$ into equation (2) we obtain the estimate that in large asexual or random mating sexual populations there should be on average 100 deleterious mutations per individual at equilibrium. In comparison, consider a population in which females choose the 'best of 4' mates, which seems a reasonable number for birds at least^{21,22}. In this case, the average number of deleterious mutations per individual will be only 31 if the number of mutations borne by an individual explains 6% of the variance in the chosen trait t , which gives a reproductive advantage to the sexual population over an asexual population of about two, and hence accounts for the maintenance of sex. If, on the other hand, only 1% of the variance in the chosen trait is explained by the number of deleterious mutations—and such a low amount seems unlikely as females seem to be quite accurate in distinguishing between phenotypes²²—the average number of mutations would be reduced to 60, which would give the choosy sexual population an advantage of 1.5 over an otherwise equivalent asexual population. Although not enough to explain sex on its own, mate choice clearly may contribute significantly to a pluralistic explanation of sex⁶, in this case leaving a factor of only 1.33 to be explained by other causes.

In general, the twofold cost of sex will be just balanced by mate choice if it reduces the average number of mutations by $\ln 2/hs$. Equation (4) may be rearranged to give the critical value of ρ_{mt} (or other parameters) in terms of the remaining parameters. Thus:

$$\rho_{mt} = \frac{2(\ln 2)(hs)^{1/2}}{\alpha_n(U - \ln 2)^{1/2}} \quad (6)$$

Substitution of reasonable parameter values into this equation shows that the critical ρ_{mt} will be small provided U is not too close to $\ln 2$, the usual limiting value for mutational explanations of sex (see Fig. 1).

Although best of n sexual selection leads to a synergistic, epistatic fitness function for males, the mechanism described above does not require it, and in this way it differs significantly from Kondrashov's mutational deterministic hypothesis⁵. The mechanism described

here works for mate choice that leads to a multiplicative fitness component to male fitness, or even when mutations satisfy antagonistic epistasis. To see this all one need do is replace the best of n mate choice term in equation (3) with any increasing function of mutation number. In particular, suppose that mutations have a multiplicative marginal effect of $h's'$ per mutation on male mating success (i.e. no epistasis). The term in equation (3) owing to best of n mate choice would then be replaced by the term $-\frac{1}{2}h's'm_i$. This leads to an equilibrium number of deleterious mutations (compare equations (2) and (4)):

$$\bar{m} \approx \frac{U}{hs + \frac{1}{2}h's'} \quad (7)$$

So, in this case, if variation in mating success accounts for two-thirds of the genetically determined variance in a males reproductive output, then the genetic viability load of the population would be halved.

There is another way of looking at this: the selection gradient experienced by an allele in a sexual population is the average of the selection gradients it experiences in males and females, but the reproductive output of an anisogamous sexual lineage relative to an otherwise equivalent asexual lineage is determined solely by the relative reproductive outputs of females in the two populations. If males undergo harsher selection than sexual and asexual females, as evinced by their generally higher variance in lifetime reproductive success, then deleterious alleles undergo harsher selection in a sexual lineage than in an asexual one and thus will have their frequency reduced. Asexual females, therefore, will bear, on average, more deleterious mutations than sexual ones.

In Kondrashov's theory² the role of sex is to increase the variance in the distribution of mutations post-selection relative to an equivalent asexual population so that synergistic epistasis can act more efficiently in stripping out deleterious alleles. The mechanism presented above has the added feature of exchanging the viability load of the females in the population for the soft load of differential male mating success, as can be seen from the form of equations (4) and (7) compared to equation (2). Differential male mating success is equivalent, in terms of removing deleterious mutations, to increased male mortality. Males are a sunk cost with no equivalent in an asexual lineage arising from a sexual one. What happens to

them in terms of mating success (or survival) does not impinge on the total reproductive output of the sexual line (other than the mutation reduction under discussion). Mate choice improves the reproductive output of a sexual female lineage relative to an asexual or random-mating sexual female lineage by transferring viability load to differential male mating success.

Viewing the right side of equation (3) as the average of male and female selection gradients highlights a number of caveats with respect to the hypothesis presented here. A mutation that is beneficial to males but deleterious to females will have greater representation in a sexual population than in an asexual one as it will undergo reduced selection and, if its benefit to males outweighs its disadvantage to females, the allele will spread to fixation. In asexual females such a mutation will always encounter opposing selection and be kept at low level. In this way sex may reduce the base reproductive output of sexual females below that of asexual females. For example, the experimental removal of sexual selection in *Drosophila melanogaster* removes a reproductive load due to antagonistic coevolution between the sexes²³. Furthermore, the average mutation rate in a metazoan sexual population may be greater than in an equivalent asexual population owing to the generally greater number of germline mitotic divisions in males. In this instance sexual populations are also at a prima facie disadvantage to asexual ones. On the other hand, if males undergo greater viability selection than females, then the average viability selection gradient hs will be greater in a sexual population than in an asexual one and thus sexual females would bear fewer mutations than asexual ones, even with random mating.

As with other mutational theories, the significance of these results with respect to the maintenance of sex rests on genomic mutation rates, U . Clearly, from equation (1), for the full twofold cost of sex to be paid for by reduced mutational loads, U has to be at least $\ln 2 \approx 0.7$. The mutation rate for organisms in general has been a matter of much recent debate, with different studies coming to widely varying estimates^{20,24–27}.

All that is required for the functioning of the above model is that there is a correlation between the number of deleterious alleles and mating success; mate choice need not be for 'good genes' per se. For example, females might choose males on the basis of the quality of their territory because of the direct benefits they obtain, but this

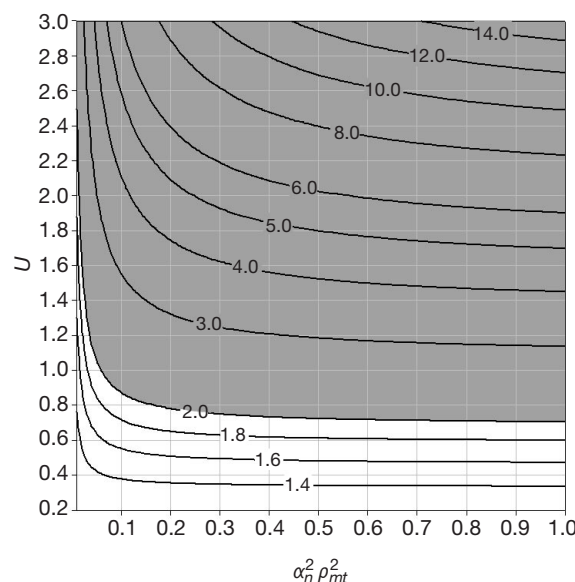


Figure 1 Contour graph of the relative viability advantage of a sexual population with best of n mate choice over an equivalent asexual population $[(1 - L_{\text{sex}})/(1 - L_{\text{asex}})]$. U represents the per generation genomic deleterious mutation rate, and ρ_{mt} is the

correlation between the number of deleterious mutations carried by males and the female rank ordering of males. The full twofold cost of sex is paid by reduction of load in the shaded region.

does not preclude male territory quality being correlated with the number of deleterious alleles they bear and, as can be seen from Fig. 1 and equation (6), strong effects on load do not require strong correlations. Although the calculations that have been presented are based on best of n mate choice, mathematically the results are much more general. Any form of differential mating success that leads to the average father bearing fewer deleterious mutations than the average male will have similar effects.

Although sexual selection and differential male mating success is normally discussed in relation to animals, there is no reason why the mechanism outlined here cannot operate in plants. The preconditions are certainly satisfied²⁸.

Finally, from the perspective of viability, sexual selection often leads to the evolution of poorly adapted traits and destructive competition for mates. These are not a further cost of sex. Any asexual lineage arising from a sexual population would be very unlikely to lose costly female secondary sexual characteristics at the same time as becoming parthenogenetic. Moreover, males are already a sunk cost included in the twofold cost of sex so, unlike females, population fitness is unaffected by what happens to them individually. As illustrated above, this destructive competition and extravagant ornamental waste actually reduces the cost of sex by reducing the population's mutational genetic load. Consequently, sexual selection reduces mutational constraints on the size of genomes and the number of germline mitotic divisions. The fact that sexual organisms began to choose their mates may have had far-reaching consequences for the evolution of complex life. □

Received 20 October 2000; accepted 5 April 2001.

- Maynard Smith, J. in *Group Selection* (ed. Williams, G. C.) 163–175 (Aldine–Atherton, Chicago, 1971).
- Kondrashov, A. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).
- Kondrashov, A. Classification of hypotheses on the advantages of amphimixis. *J. Hered.* **84**, 372–387 (1993).
- Hurst, L. D. & Peck, J. R. Recent advances in understanding of the evolution and maintenance of sex. *Trends Evol. Ecol.* **11**, 46–52 (1996).
- Barton, N. H. & Charlesworth, B. Why sex and recombination? *Science* **281**, 1986–1989 (1998).
- West, S. A., Lively, C. M. & Read, A. F. A pluralist approach to sex and recombination. *J. Evol. Biol.* **12**, 1003–1012 (1999).
- Lenski, R. E., Ofria, C., Collier, T. C. & Adams, C. Genome complexity, robustness and genetic interactions in digital organisms. *Nature* **400**, 661–664 (1999).
- Elena, S. F. & Lenski, R. E. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
- De Visser, J. A. G. M. & Hoekstra, R. F. Synergistic epistasis between loci affecting fitness: evidence in plants and fungi. *Genet. Res. Camb.* **71**, 39–49 (1998).
- West, S. A., Peters, A. D. & Barton, N. H. Testing for epistasis between deleterious mutations. *Genetics* **149**, 435–444 (1998).
- Bell, G. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality* (Univ. California Press, Berkeley, 1982).
- Trivers, R. L. Sexual selection and resource-accruing abilities in *Anolis garmani*. *Evolution* **30**, 253–269 (1976).
- Atmar, W. On the role of males. *Anim. Behav.* **41**, 195–205 (1991).
- Koeslag, P. D. & Koeslag, J. H. Koinophilia stabilizes bi-gender sexual reproduction against asexual in an unchanging environment. *J. Theor. Biol.* **166**, 251–260 (1994).
- Kodric-Brown, A. & Brown, J. H. Anisogamy, sexual selection, and the evolution and maintenance of sex. *Evol. Ecol.* **1**, 95–105 (1987).
- Manning, J. T. Males and the advantage of sex. *J. Theor. Biol.* **108**, 215–220 (1984).
- Kondrashov, A. S. & Crow, J. F. King's formula for the mutation load with epistasis. *Genetics* **120**, 853–856 (1988).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).
- Rohlf, F. J. & Sokal, R. R. *Statistical Tables* (W. H. Freeman, New York, 1994).
- Simmons, M. J. & Crow, J. F. Mutations affecting fitness in *Drosophila* populations. *Ann. Rev. Genet.* **11**, 49–78 (1977).
- Petrie, M., Halliday, T. & Sanders, C. Peahens prefer peacocks with elaborate trains. *Anim. Behav.* **41**, 323–332 (1991).
- Dale, S., Rinde, H. & Slagsvold, T. Competition for a mate restricts mate search of female pied flycatchers. *Behav. Ecol. Sociobiol.* **30**, 165–176 (1992).
- Holland, B. & Rice, W. R. Experimental removal of sexual selection reverses intersexual antagonistic coevolution and removes a reproductive load. *Proc. Natl Acad. Sci. USA* **96**, 5083–5088 (1999).
- Lynch, M. et al. Perspective: spontaneous deleterious mutation. *Evolution* **63**, 645–663 (1999).
- Keightley, P. D. & Eyre-Walker, A. Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**, 515–523 (1999).
- Keightley, P. D. & Eyre-Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331–333 (2000).
- Kondrashov, A. S. Sex and U. *Trends Genet.* **17**, 75–77 (2000).
- Arnold, S. J. et al. Sexual selection in plants and animals. *Am. Nat.* **144**, S1–S149 (1994).

Acknowledgements

I would like to thank R. Dawkins, L. Eshel, A. Grafen, L. Hurst, A. Kacelnik, A. Kondrashov, J. Metcalf, R. Nair, S. Schultz, F. Weissing, and in particular, M. Ridley for their comments on drafts of this manuscript.

Correspondence and requests for materials should be addressed to S.S. (e-mail: steven.siller@zoo.ox.ac.uk).

Sexual selection and the maintenance of sexual reproduction

Aniel F. Agrawal

Department of Biology, Indiana University, Bloomington, Indiana 47405-3700, USA

The maintenance of sexual reproduction is a problem in evolutionary theory because, all else being equal, asexual populations have a twofold fitness advantage over their sexual counterparts^{1,2} and should rapidly outnumber a sexual population because every individual has the potential to reproduce. The twofold cost of sex exists because of anisogamy or gamete dimorphism²—egg-producing females make a larger contribution to the zygote compared with the small contribution made by the sperm of males, but both males and females contribute 50% of the genes. Anisogamy also generates the conditions for sexual selection³, a powerful evolutionary force that does not exist in asexual populations. The continued prevalence of sexual reproduction indicates that the ‘all else being equal’ assumption is incorrect. Here I show that sexual selection can mitigate or even eliminate the cost of sex. If sexual selection causes deleterious mutations to be more deleterious in males than females, then deleterious mutations are maintained at lower equilibrium frequency in sexual populations relative to asexual populations. The fitness of sexual females is higher than asexuals because there is no difference in the fecundity of sexual females and asexuals of the same genotype, but the equilibrium frequency of deleterious mutations is lower in sexual populations. The results are not altered by synergistic epistasis in males.

Sexual reproduction with anisogamy generates the conditions for sexual selection³. “Competition for mates usually characterizes males because males usually invest almost nothing in their offspring” (ref. 4). Sexual selection on males (through male–male competition and/or female choice) is realized as the variance in male mating success; this variance can be large, making sexual selection very strong⁵. Sexual selection is a powerful force responsible for driving marked morphological, physiological, and behavioural change in many species³. Sexual selection occurs in sexual populations, but does not affect asexual ones. This fundamental difference between sexual and asexual populations has been largely ignored (but see refs 6–8).

Several prominent theories for the evolution of sexual reproduction focus on differences between sexual and asexual populations in the equilibrium frequency of deleterious alleles at mutation–selection balance^{9–13}. Because of recurrent mutation, neither sexual nor asexual populations can ever be completely free of deleterious mutations and, as a result, both suffer from mutation load¹⁴. Under some conditions, sexual reproduction allows mutations to be cleared more efficiently relative to asexual reproduction, and thus sexual populations can suffer substantially less mutation load than their asexual counterparts. These theories require specific forms of gene action such as epistasis^{9–12} or dominance¹³. There is little empirical support for synergistic epistasis¹⁵. Although there is

considerable support for dominance¹⁶, it is unlikely that deleterious alleles are sufficiently recessive to eliminate the cost of sex^{13,17}. The impact of sexual selection on mutation load has received little attention even though sexual selection can reduce the mutation load contributed by a single locus⁶.

Here I consider the effect of sexual selection on mutation load by applying different fitness functions to females and males. The fitness of a female (equation (1a)) and a male (equation (1b)) carrying i deleterious mutations is:

$$w_f(i) = \exp - (a_f i + b_f i^2) \quad (1a)$$

$$w_m(i) = \exp - (a_m i + b_m i^2) \quad (1b)$$

Synergistic epistasis¹² occurs when $b > 0$. When $b = 0$, there is no epistasis and the effects of loci are independent (that is, multiplicative effects on fitness).

Because synergistic epistasis can generate an advantage to sexual reproduction^{10–12} even without sexual selection, I first consider the case where there is no epistasis, $b_f = b_m = 0$. With $a_m = \alpha a_f$, values of α greater than one indicate that sexual selection exacerbates the effects of deleterious alleles in males. When there is no epistasis, the equilibrium mean fitness of females can be solved analytically (Box 1):

$$W_{\text{SEX}} = \exp \left[\frac{4M(e^{-a_f} - 1)}{2 - e^{-a_f} - e^{-\alpha a_f}} \right] \quad (2)$$

where M is the average number of mutations per gamete per generation.

To determine when mutation load provides an advantage to sexual reproduction, I consider the ratio of mean fitness of sexual females to asexuals, $R = W_{\text{SEX}}/W_{\text{ASEX}}$. Regardless of the form of selection^{9,10}, $W_{\text{ASEX}} = e^{-2M}$. In the absence of sexual selection, ($\alpha = 1$), equation (2) reduces to $W_{\text{SEX}} = e^{-2M} = W_{\text{ASEX}}$. Thus, the standard result¹² is recovered: without sexual selection, there is no advantage to sex under multiplicative selection (that is, $R = 1$).

With sexual selection ($\alpha > 1$), R is greater than one. When $R \geq 2$, the reduced mutation load of sexual populations completely compensates for the twofold cost of sex. The maximum possible value of R occurs in the hypothetical situation of no mutation load in sexual populations ($W_{\text{SEX}} = 1$), in this case $R_{\text{MAX}} = e^{2M}$. Thus R_{MAX} can only be greater than or equal to two when $M > 0.346$. This

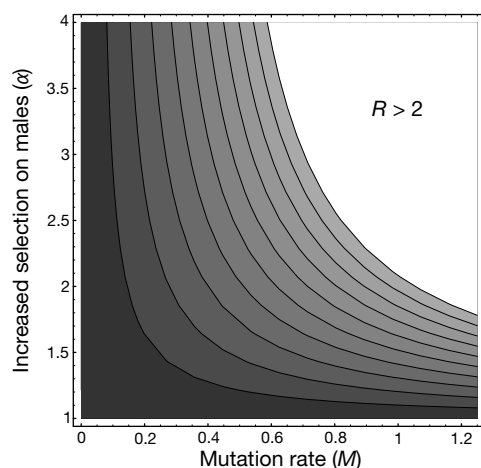


Figure 1 Effects of mutation rate and sexual selection on the cost of sex. A contour plot shows the mean fitness of sexual females relative to asexual individuals (R) as a function of the mutation rate (M) and the degree to which selection is stronger on males than females (α). The plot is generated using equation (2) with $a_f = 0.02$. Contour lines of $R = 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9$ and 2 are shown. R values increase from dark to light shading.

mutation rate sets a lower bound required for any deterministic mutation-based hypothesis^{9–13} to completely explain sex. Nonetheless, at lower mutation rates, mutation-based hypotheses can pay part of the cost of sex. Any value of R greater than one indicates that a sexual population has an advantage over an asexual population with respect to mutation load. Values of R are plotted in Fig. 1. The advantage to sex increases with both α and M .

Synergistic epistasis for fitness may be a common outcome of competition for mates¹⁸. Does synergistic epistasis in males alter the advantage of sex? It is possible to compare the advantage to sex with multiplicative selection males ($a_m \neq 0, b_m = 0$) to the advantage with synergistically epistatic selection males ($a_m = 0, b_m \neq 0$) by calculating the average selection coefficient against deleterious alleles (Box 1). When multiplicative and epistatic selection are of similar average strengths, both forms of selection produce similar R values (Fig. 2). Thus, synergistic epistasis in males alone does not provide an advantage to sex unless the average selection is stronger in males than in females. Notably, synergistic epistasis in females alone does provide an advantage to sex (Table 1).

For sexual selection to maintain sexual reproduction against asexuality, the mutation rate must be on the order of one, and selection against deleterious alleles must be about twice as strong in males as in females (Figs 1 and 2). Although there is consensus that mutation rates are much less than unity for prokaryotes and greater than unity for long-lived animals such as humans, the mutation rates for most eukaryotes are hotly contested^{19,20}. Some studies²⁰ report values of the order of one but others argue that it is at least one order of magnitude lower^{19,21}.

There are no estimates of the extent to which selection is stronger on males than females. There is, however, no doubt that sexual selection on males is a powerful and ubiquitous force in sexual populations³. An analysis of selection studies suggests that sexual selection is usually stronger than natural selection²². Although studies of sexual selection tend to focus on gaudy display characters in males, sexual selection may well affect the entire genome⁷. Only a relatively small fraction of genes are responsible for how a male allocates his resources to different characteristics (including sexually selected characteristics). In contrast, most genes in the genome are involved, in some way, with the acquisition of resources. On average, a male with more resources will be better able to compete for females than a male with less resources.

Existing data shows that males that obtain more mating success have better than average genomes. For example, in house mice, offspring viability and performance is higher when mothers are allowed to choose their mates than when they are forced to mate randomly²³. In a variety of other species, male secondary sexual

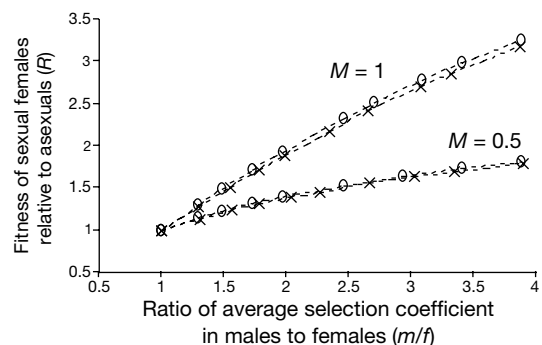


Figure 2 Multiplicative versus synergistically epistatic effects in males. The mean fitness of sexual females relative to asexual individuals (R) is plotted as a function of the average selection coefficient in males (m) relative to the average selection coefficient in females (f) (see Box 1). Circles, values under multiplicative fitness effects in males ($a_m \neq 0, b_m = 0$); crosses, values under epistatic effects in males ($a_m = 0, b_m > 0$). These values were generated using computer simulations with $a_f = 0.02$ and $b_f = 0$.

Box 1

The main mathematical details of the model

The multilocus approach in ref. 9 is followed closely. Unlike that study, I use an exponential fitness function¹² (equation (1)) and allow for different selection on males and females.

The analytical solution presented below assumes no epistasis ($b_i = b_m = 0$). Let p_i be the frequency of individuals with i mutations before selection. The average fitness of females (with respect to mutation load) is $W_f = \sum_{i=0}^{\infty} p_i w_f(i)$ and the average fitness of males is $W_m = \sum_{i=0}^{\infty} p_i w_m(i)$. The average selection coefficient⁹ against each mutation in females (equation (3a)) and males (equation (3b)) is:

$$f = \sum_{i=0}^{\infty} p_i (w_f(i+1) - w_f(i)) / W_f \quad (3a)$$

$$m = \sum_{i=0}^{\infty} p_i (w_m(i+1) - w_m(i)) / W_m \quad (3b)$$

Because of sexual selection, the mean number of mutations in males after selection, λ'_m , is different from the mean number of mutations in females after selection, λ'_f . Because of free recombination, I assume that the number of mutations in male gametes and the number of mutations in female gametes will both follow Poisson distributions, but that they will have different means. Zygotes are produced by the fusion of randomly chosen male and female gametes. The probability of a zygote with i mutations is:

$$p_i = \sum_{x=0}^i \left(\frac{(0.5\lambda'_f + M)^x}{x!} e^{-(0.5\lambda'_f + M)} \right) \left(\frac{(0.5\lambda'_m + M)^{i-x}}{(i-x)!} e^{-(0.5\lambda'_m + M)} \right) \quad (4a)$$

$$= (e^{-0.5(\lambda'_m + \lambda'_f) + 2M}) \sum_{x=0}^i \left(\frac{(0.5\lambda'_m + M)^x (0.5\lambda'_f + M)^{i-x}}{x!(i-x)!} \right) \quad (4b)$$

$$= \frac{(0.5(\lambda'_m + \lambda'_f) + 2M)^i}{i!} e^{-(0.5(\lambda'_m + \lambda'_f) + 2M)} \quad (4c)$$

This equation shows that the number of mutations in each zygote of the next generation is Poisson-distributed with mean $\lambda = 0.5(\lambda'_m + \lambda'_f) + 2M$.

Using the Poisson distribution and setting $a_m = \alpha a_f$, I find:

$$W_f = \exp[-\lambda(1 - e^{-a_f})] \quad (5a)$$

$$W_m = \exp[-\lambda(1 - e^{-a_m})] \quad (5b)$$

The average selection coefficients (equation (3)) are $f = 1 - e^{-a_f}$ and $m = 1 - e^{-a_m}$.

Censusing after mutation and before selection, the mean number of mutations in the next generation is

$$\lambda(t+1) = 2M + 0.5(e^{-a_f} + e^{-a_m})\lambda(t) \quad (6)$$

At equilibrium there is no change to the mean number of mutations, so the equilibrium mean number of mutations, λ_{eq} , can be found by setting $\lambda(t+1) = \lambda(t)$ in equation (6). I find $\lambda_{eq} = 4M/(2 - e^{-a_f} - e^{-a_m})$. Substituting this λ_{eq} into equation (5a) yields the equilibrium fitness of females shown in equation (2).

These analytical solutions produce the same results as those found using deterministic computer simulations (as described in ref. 9). Briefly, zygotes are formed by the fusion of a randomly chosen male and female gamete. Selection is applied to females (equation (1a)) and males (equation (1b)). Females (or males) carrying different numbers of mutations produce eggs (or sperm) in proportion to their frequency after selection. Gametes are produced assuming free recombination; an individual with i mutant alleles produces gametes with various numbers of mutant alleles after a binomial distribution with i trials and a probability of success of 0.5. Equilibrium was usually reached in a few hundred generations (less than 500). Results of computer simulations were consistent with those reported in refs 10 and 13. All results involving epistasis (Fig. 2 and Table 1) were found using computer simulations.

Table 1 Synergistic epistasis in females versus males

Form of selection		Fitness of sexuals relative to asexuals (F)
Female	Male	
Mult	Mult	1.00
Mult	Epi	0.99
Epi	Mult	2.65
Epi	Epi	2.63

The effect of selection of different forms (but of equal strength) on females and males is shown with a representative example. Values were generated by computer simulations with $M = 1$. Multiplicative selection (Mult) was produced with $a = 0.02$ and $b = 0$; synergistically epistatic selection (Epi) was produced with $a = 0$ and $b = 0.0002$. These different forms of selection produce roughly equally strong average selection coefficients at equilibrium, $m/f \approx 1$. Synergistic epistasis only provides an advantage to sex when it occurs in females.

characteristics used in female choice are reliable predictors of offspring fitness (see ref. 24 and references therein). Even sperm competition, a cryptic form of sexual selection, results in the production of higher quality offspring²⁵. Similarly, pollen competition in plants can result in the production of higher performance seedlings²⁶. One study on *Drosophila melanogaster* found that a substantial portion of the selection against some mutations results from differential mating success²⁷.

Cases in which sexual selection is weak because males provide paternal care^{3,4} do not challenge the theory presented here, as the twofold cost of sex is only expected if males contribute nothing beyond genes to their offspring². One review notes that asexual reproduction is more common in isogamous than anisogamous lineages even though the cost of sex is expected to be greater in the latter²⁸. I have shown that sexual selection on males (a result of anisogamy) can help to purge deleterious alleles from the entire sexual population. In this way, males can reduce and even eliminate the twofold cost of sex. In addition to removing deleterious alleles, theoretical work indicates that sexual selection can also increase the probability of fixation of new, beneficial mutations²⁹. When the effects of sexual selection are combined to other advantages of sexual reproduction³⁰, the cost of sex could easily be overcome. □

Received 9 October 2000; accepted 19 March 2001.

- Maynard-Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, Cambridge, 1978).
- Bell, G. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality* (Univ. of California Press, Berkeley, 1982).
- Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, 1994).
- Trivers, R. L. in *Sexual Selection and the Descent of Man* (ed. Campbell, B. G.) 141 (Aldine, Chicago, 1972).
- Arnold, S. J. & Wade, M. J. On the measurement of natural and sexual selection: applications. *Evolution* **38**, 720–734 (1984).
- Manning, J. T. Males and the advantage of sex. *J. Theor. Biol.* **108**, 215–220 (1984).
- Kodric-Brown, A. & Brown, J. H. Anisogamy, sexual selection, and the evolution and maintenance of sex. *Evol. Ecol.* **1**, 95–105 (1987).
- Koselag, J. H. & Koeslag, P. D. Evolutionary stable meiotic sex. *J. Hered.* **84**, 396–399 (1993).
- Kimura, M. & Maruyama, T. The mutational load with epistatic interactions in fitness. *Genetics* **54**, 1337–1351 (1966).
- Kondrashov, A. S. Selection against harmful mutations in large sexual and asexual populations. *Genet. Res.* **40**, 325–332 (1982).
- Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).
- Charlesworth, B. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.* **55**, 199–221 (1990).
- Chasnov, J. R. Mutation-selection balance, dominance and the maintenance of sex. *Genetics* **156**, 1419–1425 (2000).
- Crow, J. F. in *Mathematical Topics in Population Genetics* (ed. Kojima, K.-I.) 128–177 (Springer, Berlin, 1970).
- Elena, S. E. & Lenski, R. E. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Sunderland, 1998).
- García-Dorado, A. & Caballero, A. On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* **155**, 1991–2001 (2000).
- Peck, J. R. & Waxman, D. Mutation and sex in a competitive world. *Nature* **406**, 399–404 (2000).
- Keightley, P. D. & Eyre-Walker, A. Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**, 515–523 (1999).
- Lynch, M. et al. Spontaneous deleterious mutation. *Evolution* **53**, 645–663 (1999).
- Keightley, P. D. & Eyre-Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331–333 (2000).
- Kingsolver, J. G. et al. The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245–261 (2001).

23. Drickamer, L. C., Gowaty, P. A. & Holmes, C. M. Free female choice in house mice affects reproductive success and offspring viability and performance. *Anim. Behav.* **59**, 371–378 (2000).
24. Welch, A. M., Semlitsch, R. D. & Gerhardt, H. C. Call duration as an indicator of genetic quality in male gray tree frogs. *Science* **280**, 1928–1930 (1998).
25. Evans, J. P. & Magurran, A. E. Multiple benefits of multiple mating in guppies. *Proc. Natl Acad. Sci. USA* **97**, 10074–10076 (2000).
26. Mulcahy, D. L. The rise of angiosperms: a genecological factor. *Science* **206**, 20–23 (1979).
27. Whitlock, M. C. & Bourget, D. Factors affecting the genetic load in *Drosophila*: synergistic epistasis and correlations among fitness components. *Evolution* **54**, 1654–1660 (2000).
28. Hurst, L. D. & Peck, J. R. Recent advances in understanding of the evolution and maintenance of sex. *Trends Ecol. Evol.* **11**, 46–52 (1996).
29. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**, 1855–1861 (2000).
30. Howard, R. S. & Lively, C. M. Parasitism, mutation accumulation and the maintenance of sex. *Nature* **367**, 554–557 (1994).

Acknowledgements

I thank C. Lively, K. Quinlan and M. Wade for comments. A. Kondrashov suggested using the multilocus approach to this problem. S. Otto pointed out that the analytical solution is exact by use of equation (6). This work was supported by NSERC Canada.

Correspondence and requests for materials should be addressed to A.E.A. (e-mail: aagrawal@bio.indiana.edu).

PCR amplification of the Irish potato famine pathogen from historic specimens

Jean B. Ristaino*, Carol T. Groves*† & Gregory R. Parra*

* Department of Plant Pathology, North Carolina State University, Raleigh, North Carolina 27695, USA

Late blight, caused by the oomycete plant pathogen *Phytophthora infestans*, is a devastating disease of potato and was responsible for epidemics that led to the Irish potato famine in 1845 (refs 1–5). Before the 1980s, worldwide populations of *P. infestans* were dominated by a single clonal lineage, the US-1 genotype or Ib mitochondrial DNA (mtDNA) haplotype, and sexual reproduction was not documented outside Mexico, the centre of diversity of the pathogen^{6,7}. Here we describe the amplification and sequencing of 100-base-pair fragments of DNA from the internal transcribed spacer region 2 from 28 historic herbarium samples including Irish and British samples collected between 1845 and 1847, confirming the identity of the pathogen. We amplified a variable region of mtDNA that is present in modern Ib haplotypes of *P. infestans*, but absent in the other known modern haplotypes (Ia, IIa and IIb)⁸. Lesions in samples tested were not caused by the Ib haplotype of *P. infestans*, and so theories that assume that the Ib haplotype is the ancestral strain need to be re-evaluated^{4,7}. Our data emphasize the importance of using historic specimens when making inferences about historic populations.

During the nineteenth and early twentieth centuries, scientists collected and preserved potato and tomato leaves infected with *P. infestans* from the Irish potato famine, and these specimens can be used today to answer questions about the population biology of the pathogen⁹ (Fig. 1). Disputes about nomenclature, phylogenetics, function and evolution of genes, and origins of populations can be addressed using herbarium collections^{10–17}. We used polymerase chain reaction (PCR) amplification and sequencing of ribosomal DNA (rDNA) and mtDNA of *P. infestans* from herbarium speci-

mens to determine whether the Ib mtDNA haplotype was the clonal lineage responsible for historic epidemics.

Dried potato and tomato lesions sampled from the Royal Botanic Gardens Mycological Herbarium, Kew, UK, and the US National Fungus Collection, Beltsville, Maryland, were aseptically removed from specimen envelopes and placed in sterile microfuge tubes. Initial DNA extractions from British and European samples were conducted in the Jodrell Laboratory at Kew in a laboratory without a history of work with the pathogen. Non-infected potato leaves and leaves infected with *P. infestans* from modern epidemics were pressed and dried in the Mycological Herbarium in the Department of Plant Pathology, North Carolina State University. All manipulations of DNA from modern samples were performed in our laboratory in the Department of Plant Pathology. Work with the herbarium specimens was performed in the North Carolina State University Phytotron Containment Facility Laboratory, which was equipped with separate supplies, reagents and equipment and has no history of research involving *P. infestans*.

DNA was extracted from herbarium samples according to a modification of a cetyltrimethylammonium bromide (CTAB) procedure¹⁸. DNA was diluted 1:10 and used with PCR primers PINF and HERB1 and PINF and ITS3 from 123 herbarium samples. Thirty-nine samples (31%) yielded the expected product (around 100 base pairs (bp) in size) when amplified with primers PINF and HERB1 (see Supplementary Information). Only three samples (2.4%), all from the 20th century, yielded a product approximately 300 bp in size when amplified with primers PINF and ITS3. These results are expected because DNA tends to degrade in ancient materials and primers that amplify smaller product sizes (<200 bp) are more effective in these types of specimen¹³. DNA of *P. infestans* was amplified from 39 samples, including four samples of *Solanum tuberosum* collected in Britain and France in 1845, four samples collected in Ireland and Britain in 1846, and one sample collected in Britain in 1847 (Fig. 2; and see Supplementary Information). We also amplified DNA of *P. infestans* from a sample of *Anthocercis ilicifolia*, a solanaceous evergreen shrub native to Western Australia

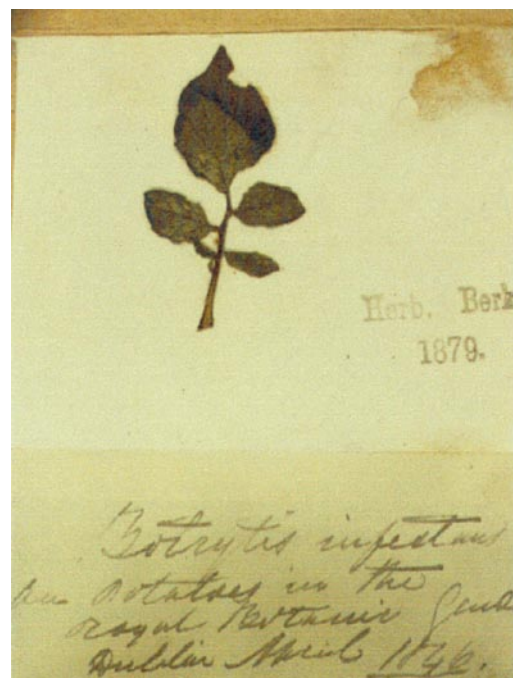


Figure 1 Specimen of potato infected with *P. infestans*. This sample was collected by J. Lindley in 1846 in the Royal Botanic Gardens, Dublin, Ireland and identified by M. J. Berkeley. It is deposited in the collections at the Royal Botanic Gardens, Kew, UK and is one of the oldest known specimens of potato from the potato famine epidemics.

† Present address: USDA-ARS, New England Plant, Soil, and Water Laboratory, Orono, Maine 04469, USA.

that was introduced into the National Botanic Gardens at Glasnevin in Dublin in 1842 and collected in 1846 by David Moore. This is the earliest known definitive diagnosis of an alternative solanaceous host for *P. infestans*¹⁹.

We sequenced the DNA from internal transcribed spacer (ITS) region 2 that was amplified by the PINF and HERB1 primers from 12 of the herbarium specimens. Duplicate PCR and sequencing reactions were performed in both the forward and reverse directions to confirm the DNA sequence and to minimize error. Sequences were aligned with Clustal W, version 1.8 (ref. 20) and have been submitted to GenBank (see Supplementary Information). Analysis of the DNA sequences confirmed that the targeted region of pathogen ITS DNA was successfully amplified from the lesions. This region of ITS DNA is highly conserved among extant isolates of *P. infestans* and provides a diagnostic region for identification of the pathogen in potato tissue¹⁸. Our molecular data confirm that these historic samples were infected with *P. infestans*.

Mitochondrial haplotypes have been designated in *P. infestans* using both PCR approaches and restriction fragment length polymorphism (RFLP) analysis of mtDNA^{8,21–23}. Four haplotypes have been designated using a PCR-based approach: Ia, Ib, IIa and IIb⁸. Some of the DNA sequence mutations that generated the mtDNA polymorphisms in the four haplotypes have been identified^{8,23–25}. We have sequenced the P2 and P4 regions from three modern isolates of each of the four known mtDNA haplotypes and have confirmed the sequence variation that has been reported (see Supplementary Information)^{8,24,25}.

The mtDNA primer set P2F4 and P2R4 amplified a variable region near the 3' end of the *Nad4* gene in the P2 region of the mitochondrial genome^{8,21,24}. A 167-bp PCR product and sequence data were obtained from ten of the herbarium samples that were previously positive for the PINF and HERB1 primer set (Fig. 3a; and see Supplementary Information). Modern isolates of the Ib mtDNA haplotype contain an *MspI* restriction site in this region, whereas the other three modern haplotypes have a single base-pair change from C to T and thus lack this restriction site (Fig. 3b)^{8,23}. None of the sequences from the herbarium specimens contained the *MspI* restriction site, indicating that the lesions in the samples we examined were not caused by the Ib haplotype of *P. infestans*. It is possible that the ancestral strain belonged to the Ia haplotype, and that the larger insertion sequence found in group II haplotypes occurred more recently from a Ia haplotype²¹. We are investigating this possibility.

Isolates of *P. infestans* that display the typical RFLP fingerprint of the US-1 genotype are the Ib mtDNA haplotype^{8,21–23,26–28}. The Ib haplotype has been reported in modern isolates from Ireland, Ecuador, Brazil, the UK and the US^{21,23}. The Ia mtDNA haplotype was the dominant haplotype found in populations of *P. infestans* in Mexico, in addition to the IIa and IIb haplotypes^{22,23}. The Ib haplotype has not been found among present-day isolates of the pathogen in Mexico, even though this haplotype is proposed to be

the ancestral type^{21,23}. In studies on the origin of human populations, all major mtDNA lineages occur in Africa, and only a subset occur in the rest of the world, thus suggesting an African origin of humans²⁹. As all the known modern mtDNA haplotypes of *P. infestans* do not occur in Mexico, this might suggest an alternative centre of origin for the pathogen, perhaps in South America, which is the ancestral home of the potato^{1,2}. Mexico is clearly a centre of diversity for the late blight pathogen^{4,5}, but the pathogen's centre of diversity may not coincide with its centre of origin.

This is the first time, to our knowledge, that ITS DNA and mtDNA of a plant pathogen have been amplified and sequenced from lesions of samples from historic epidemics. Sequence data from the mtDNA indicated that the Ib haplotype did not cause the

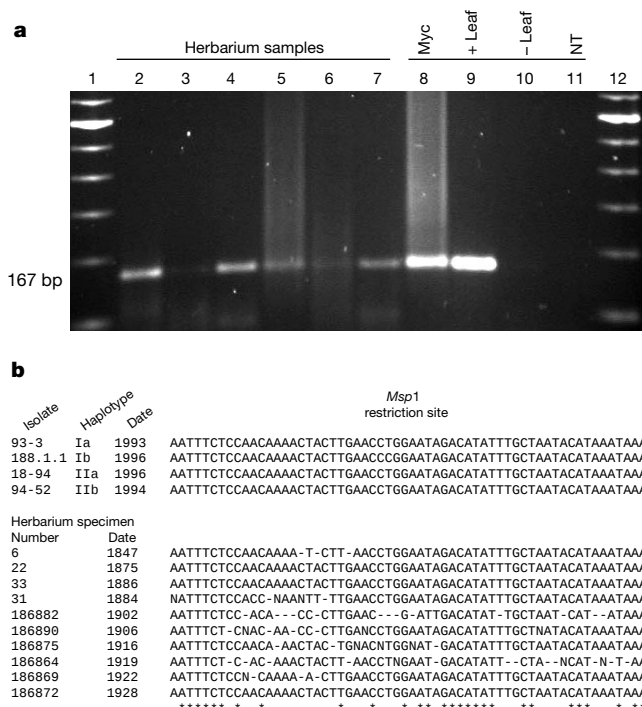


Figure 3 Amplified PCR products (P2F4/P2R4) from a variable region near the 3' end of the *Nad4* gene in the P2 region of mtDNA from herbarium specimens infected with *P. infestans*. **a**, Herbarium specimens from England: 22, Vize, 1875; 31, Grove, 1884; 33, Vize, 1886 (lanes 2–4); and from Connecticut: 186875, 186869 and 186872 collected by G. P. Clinton between 1902 and 1928 (lanes 5–7); positive controls: mycelium of *P. infestans* (Myc) and dried infected modern lesion (+ leaf); negative controls: dried non-infected leaf (– leaf) and non-template (NT) control (lanes 8–11); 100-bp ladders in lanes 1 and 12. **b**, MtDNA sequences around the *MspI* restriction site in P2 region. This restriction site is present in the Ib haplotype and absent in other extant haplotypes. See Supplementary Information and GenBank for full sequences.

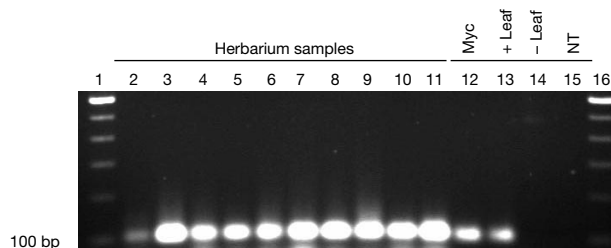


Figure 2 Amplified PCR products obtained using HERB1 and PINF from historic herbarium samples infected with *P. infestans* and collected between 1845 and 1886. PCR products from specimen: number 16, 71, 7, 47, 6, 53, 24, 22, 43, 33 (lanes 2–11); positive controls from mycelium of *P. infestans* (Myc), dried infected modern lesion (+ leaf) and

negative controls from dried non-infected leaf (– leaf) and non-template (NT) control (lanes 12–15); 100-bp DNA molecular ladder (lanes 1 and 16). See Supplementary Information for specimen list and sequence data.

lesions in the samples assayed. Our data are not consistent with theories that the US-1 genotype was solely responsible for epidemics of potato late blight during the Irish potato famine and later in the 19th and early 20th centuries in the US and Europe^{4,7,23}. Work is in progress to determine whether the lesions were caused by one of the other three extant haplotypes or some unique haplotype of *P. infestans*. Herbarium collections have not been used previously to understand epidemics of the past and track pathogen migrations. In fact, many important natural history collections have recently been critically underfunded and are in danger of being lost. Our work indicates that these historic collections can be valuable to epidemiologists and population geneticists who study plant diseases. Misleading data can be derived from studies that examine only extant phylogeographic patterns of genotypes of plant pathogens. It is important to use historic specimens when making inferences about historic populations. □

Methods

DNA extraction

Small pieces (2–3 mm²) of dried tissue were placed in sterile 1.5-ml tubes, 150 µl extraction buffer (0.35 M sorbitol, 0.1 M Tris, 0.005 M EDTA, pH 7.5, 0.02 M sodium bisulphite) was added and tubes were vortexed. Nuclei lysis buffer (150 µl) containing 0.2 M Tris, 0.05 M EDTA, pH 7.5, 2.0 M NaCl and 2% CTAB (hexadecyltrimethylammonium bromide) was added followed by 60 µl of 5% sarkosyl (N-lauryl sarcosine). Tubes were vortexed then incubated at 65 °C for 15–30 min. One volume of chloroform:isoamyl alcohol (24:1) was added to each tube and centrifuged for 15 min at 13,000g at room temperature. The aqueous phase was removed to a new tube and the chloroform extraction was repeated. DNA was precipitated overnight at –20 °C in 0.1 volumes of 3 M sodium acetate, pH 8.0, and two volumes of cold 100% ethanol. The supernatant was discarded and the pellets were washed with 70% ethanol and then dried under vacuum centrifugation. DNA was resuspended in TE (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0).

PCR reactions

HERB1 (CGGACCGCTGCGAGTCC) and PINF (CTCGCTACAATAGGAGGTC) amplify a 100-bp region of rDNA in spacer region two, and PINF and ITS3 amplify an ~300-bp region of 5.8S and spacer region 2 DNA (see Supplementary Information). PCR reactions with PINF and HERB1 or PINF and ITS3 were conducted in 50-µl reaction volumes consisting of 1 µl of template DNA (1:10 dilution of original DNA extract in TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0)), 5 µl of 10× PCR buffer (100 mM Tris-HCl, 15 mM MgCl₂, 500 mM KCl, pH 8.3), 36.5 µl of sterile distilled water, 2 µl of 2.0 mM dNTPs (0.08 mM, Amersham Pharmacia), 2 µl of 10 mM MgCl₂ (0.4 mM, Sigma), and 2 µl each of a 10-mM primer (0.4 µM final concentration), and 0.4 µl of Taq (5 U µl⁻¹) DNA polymerase (2 units, Life Technologies). Cycling parameters were initial denaturation at 96 °C for 2 min, followed by 35 cycles consisting of denaturation at 96 °C for 1 min, annealing at 55 °C for 1 min, and extension at 72 °C for 2 min. A final extension at 72 °C for 10 min followed. Aerosol-protected pipette tips and negative controls (master mix and no template DNA, or non-infected dried potato leaves) were used. All manipulations of positive control DNA (modern DNA of *P. infestans* from a pure culture or modern dried lesions on potato) were performed in a lab separate from the Phytotron lab where the herbarium samples were manipulated using the same master mix reagents. Amplified products were visualized on 1.6% agarose gels run in 1× TBE (0.5 µg ml⁻¹ of ethidium bromide). All PCR reactions were repeated twice.

PCR amplifications of mtDNA with P2F4 (5'-ACCAATTGTTGCGAAAACAG-3') and P2R4 (5'-ATTACGGCGGTTTAGCACAT-3') were conducted in 30-µl reaction volumes consisting of 5 µl of template DNA (1:10 dilution of original DNA extract in TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0)), 3 µl of 10× PCR buffer (100 mM Tris-HCl, 15 mM MgCl₂, 500 mM KCl, pH 8.3), 9.77 µl of sterile distilled water, 3 µl of 2.0 mM dNTPs (200 µM) (Amersham Pharmacia), 8.25 µl of 10 mM MgCl₂ (2.75 mM) (Sigma), and 0.15 µl each of a 65-µM primer (0.325 µM), 0.48 µl of bovine serum albumin (10 mg ml⁻¹) and 0.2 µl of 5 U µl⁻¹ Taq DNA polymerase (1 U) (Life Technologies). Positive and negative controls were used as described above. Cycling parameters were initial denaturation at 94 °C for 90 s, followed by 40 cycles of denaturation at 94 °C for 40 s, annealing at 55 °C for 1 min, and extension at 72 °C for 90 s. A final extension at 72 °C for 10 min followed. Amplified products were electrophoresed on 1.6% agarose gels containing 0.5 µg ml⁻¹ of ethidium bromide with 1× TBE running buffer. All PCR reactions were repeated twice.

Received 29 November 2000; accepted 30 March 2001.

1. Berkeley, M. J. Observations, botanical and physiological on the potato murain. *J. Hort. Soc. Lond.* **1**, 9–34 (1846).
2. de Bary, M. Researches into the nature of the potato fungus—*Phytophthora infestans*. *J. R. Agr. Soc.* **2**, 239–269 (1876).
3. Gunderson, J. H., Elwood, H., Ingold, H., Kindle, A. & Sogin, M. L. Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. *Proc. Natl Acad. Sci. USA* **84**, 5823–5827 (1987).
4. Fry, W. & Goodwin, S. B. Resurgence of the Irish potato famine fungus. *Bioscience* **47**, 363–371 (1997).

5. Goodwin, S. B., Cohen, B. A., Deahl, K. L. & Fry, W. E. Migration from Northern Mexico as the probable cause of recent genetic changes in populations of *Phytophthora infestans* in the United States and Canada. *Phytopathology* **84**, 553–558 (1994).
6. Gallegly, M. E. & Galindo, J. Mating types and oospore of *Phytophthora infestans* in nature in Mexico. *Phytopathology* **48**, 274–277 (1958).
7. Goodwin, S., Cohen, B. & Fry, W. Panglobal distribution of a single clonal lineage of the Irish potato famine fungus. *Proc. Natl Acad. Sci. USA* **91**, 11591–11595 (1994).
8. Griffith, G. W. & Shaw, D. S. Polymorphisms in *Phytophthora infestans*: four mitochondrial haplotypes are detected after PCR amplification of DNA from pure cultures or from host lesions. *Appl. Env. Microbiol.* **64**, 4007–4014 (1998).
9. Ristaino, J. B. The importance of archival and herbarium materials in understanding the role of oospores in late blight epidemics of the past. *Phytopathology* **88**, 1120–1130 (1998).
10. Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O. & Raoult, D. Detection of 400-year-old *Yersinia pestis* in human dental pulp: An approach to the diagnosis of ancient septicemia. *Proc. Natl Acad. Sci. USA* **95**, 12637–12640 (1998).
11. Taylor, J. W. & Swann, E. in *Ancient DNA: Recovery and Analysis of Genetic Information from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens* (eds Herrmann, B. & Hummel, S.) 166–181 (Springer, New York, 1994).
12. Cooper, A. et al. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **404**, 704–707 (2001).
13. Paabo, S., Higuchi, R. & Wilson, A. C. Ancient DNA and the polymerase chain reaction. *J. Biol. Chem.* **264**, 9709–9712 (1989).
14. White, T. J., Bruns, T., Lee, S. & Taylor, J. in *PCR Protocols: A Guide to Methods and Applications* (eds Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J.) 315–322 (Academic, New York, 1990).
15. Poinar, G. Ancient DNA. *Am. Sci.* **87**, 446–457 (1999).
16. Krings, M. et al. Neanderthal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
17. Salo, W. L., Aufderheide, A. C., Buikstra, J. & Holcomb, T. A. Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. *Proc. Natl Acad. Sci. USA* **91**, 2091–2094 (1994).
18. Trout, C. L., Ristaino, J. B., Madritch, M. & Wangsomboondee, T. Rapid detection of *Phytophthora infestans* in late blight infected potato and tomato using PCR. *Plant Dis.* **81**, 1042–1048 (1996).
19. Erwin, D. C. & Ribiero, O. K. *Phytophthora Diseases Worldwide* (Am. Phytopathol. Soc. Press, St. Paul, MN, 1996).
20. Higgins, D. G. & Sharp, P. M. Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* **5**, 151–153 (1989).
21. Carter, D. A., Archer, S. A., Buck, K. W., Shaw, D. S. & Shattock, R. C. Restriction fragment length polymorphisms of mitochondrial DNA of *Phytophthora infestans*. *Mycol. Res.* **94**, 1123–1128 (1990).
22. Goodwin, S. B. in *Phytophthora* (eds Lucas, J. A., Shattock, R. C., Shaw, D. S. & Cooke, L. R.) 256–271 (Cambridge Univ. Press, New York, 1991).
23. Gavino, P. D. *Mitochondrial DNA evolution and its utility for population genetic studies of Phytophthora infestans*. Thesis, Cornell Univ. (1999).
24. Lang, B. F. & Forget, L. in *Genetic Maps. Locus Maps of Complex Genomes* (ed. O'Brien, S. J.) 3.133–3.135 (Cold Spring Harbor Laboratory Press, Plain View, New York, 1993).
25. Paquin, B. et al. The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. *Curr. Genet.* **31**, 320–395 (1997).
26. Day, J. P. & Shattock, R. C. Aggressiveness and other factors relating to the displacement of populations of *Phytophthora infestans* in England and Wales. *Eur. J. Plant Path.* **103**, 379–391 (1997).
27. Lebreton, L., Laurent, C. & Andrivon, D. Evolution of *Phytophthora infestans* populations in the two most important potato production areas of France during 1992–1996. *Plant Pathol.* **47**, 427–439 (1998).
28. Wangsomboondee, T., Groves, C. T., Shoemaker, P. B., Cubeata, M. A. & Ristaino, J. B. Genetic diversity of isolates of *Phytophthora infestans* from tomato and potato in North Carolina. *Phytopathology* **90**, S65 (2000).
29. Penny, D., Steel, M., Waddell, P. J. & Hendy, M. D. Improved analysis of human mtDNA sequences support a recent African origin of *Homo sapiens*. *Mol. Biol. Evol.* **12**, 863–882 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This research was supported in part by a grant from the National Geographic Society and in part by the North Carolina State University Agricultural Research Service and the North Carolina State University Office of International Programs. We thank K. May for producing the gels for final publication; J. Thomas, NCSU for use of the Phytotron Facility Laboratory; M. Chase, Jodrell Laboratory, Royal Botanic Gardens, Kew, England, for use of his lab facilities; and D. Pegler and B. Spooner, Mycology Herbarium, Royal Botanic Gardens, Kew, England; A. Rossman, US National Fungus Collections, USDA; M. Jebb, National Botanic Gardens, Glasnevin, Dublin, Ireland; and D. Pfister, Farlow Herbarium, Harvard University for providing access to specimens. Statistical advice from J. Thorne is appreciated.

Correspondence and requests for materials should be addressed to J.B.R. (e-mail: Jean_Ristaino@ncsu.edu). DNA sequence data can be found in GenBank under accession nos AF004277–AF004280 (ITS and rDNA of four modern isolates); AY027652–AY027655, AF349576–AF349587 (PINF/Herb amplification of spacer region 2 from four modern isolates and herbarium specimens); AY003908–AY003919 (mtDNA P2 region for three modern isolates of each haplotype); AF348598–AF348609 (mtDNA P4 region for three modern isolates of each haplotype); and AF349588–349601 (P2F4/R4 amplification of mtDNA from four modern haplotypes and herbarium specimens).

Retinal ganglion cells act largely as independent encoders

S. Nirenberg, S. M. Carcieri, A. L. Jacobs & P. E. Latham

Department of Neurobiology, University of California Los Angeles, 10833 Le Conte Avenue, Los Angeles, California 90095-1763, USA

Correlated firing among neurons is widespread in the visual system. Neighbouring neurons, in areas from retina to cortex, tend to fire together more often than would be expected by chance. The importance of this correlated firing for encoding visual information is unclear and controversial^{1–5}. Here we examine its importance in the retina. We present the retina with natural stimuli and record the responses of its output cells, the ganglion cells. We then use information theoretic techniques to measure the amount of information about the stimuli that can be obtained from the cells under two conditions: when their correlated firing is taken into account, and when their correlated firing is ignored. We find that more than 90% of the information about the stimuli can be obtained from the cells when their correlated firing is ignored. This indicates that ganglion cells act largely independently to encode information, which greatly simplifies the problem of decoding their activity.

A principal goal in vision research is to understand how visual stimuli are encoded in the activity of the ganglion cells, as these cells provide all the information about the visual world that the brain receives. Several studies have proposed that visual stimuli are encoded in a complex way that depends on correlated activity^{5–7}. Such activity has been described in many species, including several mammals (cat^{8–10}, rabbit¹¹ and monkey¹²). The proposal is that this activity carries information about visual stimuli that is not present in non-correlated activity^{5–7}.

If correlated firing carries information, then strategies for understanding how ganglion cells encode visual stimuli must take the correlations into account. This means that the activity of a single ganglion cell cannot be evaluated by itself, but rather must be decoded in the context of the firing patterns of other ganglion cells. If, on the other hand, correlated firing does not carry visual information, then the firing of a ganglion cell can be evaluated independently of other cells.

It has been reported^{6,7} that correlated activity can carry information, but the methods used in these reports were indirect. Here we addressed the problem directly using an information theoretic approach. We compared the amount of information that could be obtained from pairs of ganglion pairs when their correlations were taken into account with the amount of information that could be obtained from the pairs when their correlations were ignored. The extent to which information is lost when correlations are ignored is the extent to which correlations are important for encoding information.

We used the isolated mouse retina. The stimuli were natural movies, each 7 s long and repeated 300 times, and the ganglion cell responses were recorded using a multielectrode array. We used an array with closely spaced electrodes (25–100 μm apart) to ensure that we recorded from many pairs of neighbouring ganglion cells, which tend to have overlapping receptive fields and show correlated activity^{6,9,11}.

The data were screened for pairs of responses that passed two criteria. First, both responses had to be clean of contaminating spikes from other cells. This was tested by computing the autocorrelation function for each response, which gives the firing rate as a function of time relative to each spike. As neurons have a refractory period of approximately 1 ms, the autocorrelation func-

tion for a single cell should contain a central peak flanked on either side by zero firing rate for 1 ms. Non-zero firing rates in this 1-ms window reflect contamination from electronic noise or spikes from other cells. Only responses showing less than 2% contamination were used. Second, both responses had to have average firing rates above 0.5 Hz. Typical responses are shown in Fig. 1. The dataset contained 76 cells, with 5–20 cells per retina (six retinas). This gave a total of 498 cell pairs, with 10–190 pairs per retina.

We first determined the degree of correlated activity for each pair (Fig. 2a–c). This was measured as the fraction of correlated spikes produced by the pair above chance, taking into account correlations induced by the stimulus. We called this the excess correlated fraction (ECF; see Methods). The ECFs ranged from –1% to +34% (Fig. 2b), in close agreement with similar measurements reported for other mammalian species (the highest excess fraction is 27% in cat⁹ and 28% in rabbit¹¹). It is also similar to that found for pairs of cells in cat lateral geniculate nucleus (LGN), in which the highest fraction of correlated spikes is, on average, 28% (ref. 13). We then determined the timescale over which the correlations occurred (Fig. 2c). The range was from < 1 ms to 11 ms. This is similar to that observed in cat⁹ and rabbit¹¹, although it extends to slightly shorter values. Similar short timescales (< 1 ms) are found in cat LGN¹³.

To measure the amount of information the pairs of ganglion cells carried about the stimuli when their correlations were taken into account, we used standard information theoretic techniques. We treated the movie as a series of segments of fixed temporal length, with each segment regarded as a separate stimulus. We presented the movie several hundred times to generate a large set of responses (spike trains) to each segment. This allowed us to estimate the probability of getting a particular pair of responses given a particular movie segment—that is, to estimate $P(r_1, r_2|s)$, where r_1 was the response of cell 1, r_2 was the response of cell 2 and s was the movie segment. Given these conditional probabilities, we then calculated the amount of information, I , between the responses and the stimulus segments, using the standard expression¹⁵

$$I = - \sum_{r_1, r_2} P(r_1, r_2) \log_2 P(r_1, r_2) + \sum_s P(s) \sum_{r_1, r_2} P(r_1, r_2|s) \log_2 P(r_1, r_2|s) \quad (1)$$

where $P(r_1, r_2) = \sum_s P(r_1, r_2|s)P(s)$, and $P(s)$ is the probability that a given stimulus segment s occurred. The numerical method for calculating I followed ref. 16; see Methods.

To assess the role of correlations, we calculated how much

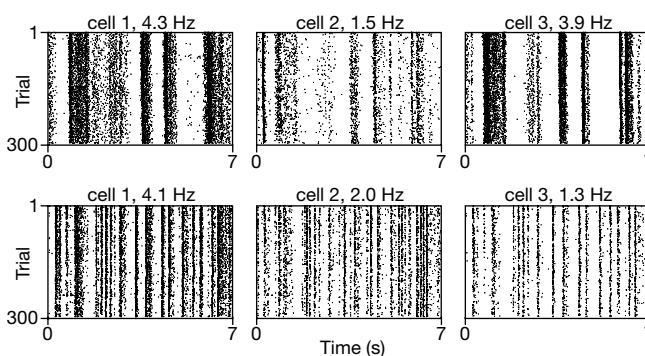


Figure 1 Typical ganglion cell responses to the movies. Each raster plot shows the response of a ganglion cell to 300 repeats of a movie. The cell's average firing rate during the movie is marked above the plot. Top, responses to Movie 1 (natural scenes); bottom, responses to Movie 3 (spatially uniform, time-varying stimulus). (See Methods for image statistics.) The top and bottom plots in each column are from the same cell.

information would be lost if the correlations in the responses of the two cells were ignored. We 'ignored' the correlations by treating the conditional probability distributions of the two responses, $P(r_1, r_2|s)$, as the product of their individual probability distributions, $P(r_1|s)$ and $P(r_2|s)$. (By definition, two responses are not correlated if their joint probability distribution equals the product of their individual ones.) We denoted this distribution $P_{\text{IND}}(r_1, r_2|s)$, with $P_{\text{IND}}(r_1, r_2|s) = P(r_1|s)P(r_2|s)$. We then used $P_{\text{IND}}(r_1, r_2|s)$ rather than the true distribution, $P(r_1, r_2|s)$, to estimate the probability of a stimulus given a response. As $P_{\text{IND}}(r_1, r_2|s)$ is not quite the true distribution, using it should lead to a loss of information. Following a method described in ref. 17 (see Methods), it can be shown that

the amount of information lost, denoted ΔI , is given by

$$\Delta I = \sum_s P(s) \sum_{r_1, r_2} P(r_1, r_2|s) \log_2 \left[\frac{P(r_1, r_2|s)}{P_{\text{IND}}(r_1, r_2|s)} \right] - \sum_{r_1, r_2} P(r_1, r_2) \log_2 \left[\frac{P(r_1, r_2)}{P_{\text{IND}}(r_1, r_2)} \right] \quad (2)$$

where ΔI is in bits, and $P_{\text{IND}}(r_1, r_2) = \sum_s P(r_1|s)P(r_2|s)P(s)$.

We found that the amount of information lost, ΔI , was remarkably small. Only one pair of cells showed a loss of information of more than 10%. Figure 3 shows the fraction of information lost for each pair, with the fraction plotted against the ECF. Although the fraction of information lost generally increased for pairs with high ECFs, the fraction lost never exceeded 11%. Thus, in the mouse retina, we find that ignoring correlations leads to only a small loss of information. Certainly, if there were pairs of ganglion cells with higher degrees of correlation, above the 34% we observed, a greater loss might be expected; however, in no mammalian species examined is the degree of correlated activity reported to be higher (as mentioned above, the highest reported ECF for cat retinal ganglion cells⁹ is 27% and for rabbit¹¹, 28%).

The loss of information was measured using responses binned at 1 ms to ensure that all information in the correlations, which occur largely on that timescale (Fig. 2c), was captured. The small bins, though, required us to use short responses, so temporal correlations beyond 10 ms were not captured. However, we also measured the loss of information using larger bins (10 ms), to better capture information in the broader correlations, and the same result was obtained: no pair showed a loss of information greater than 11%.

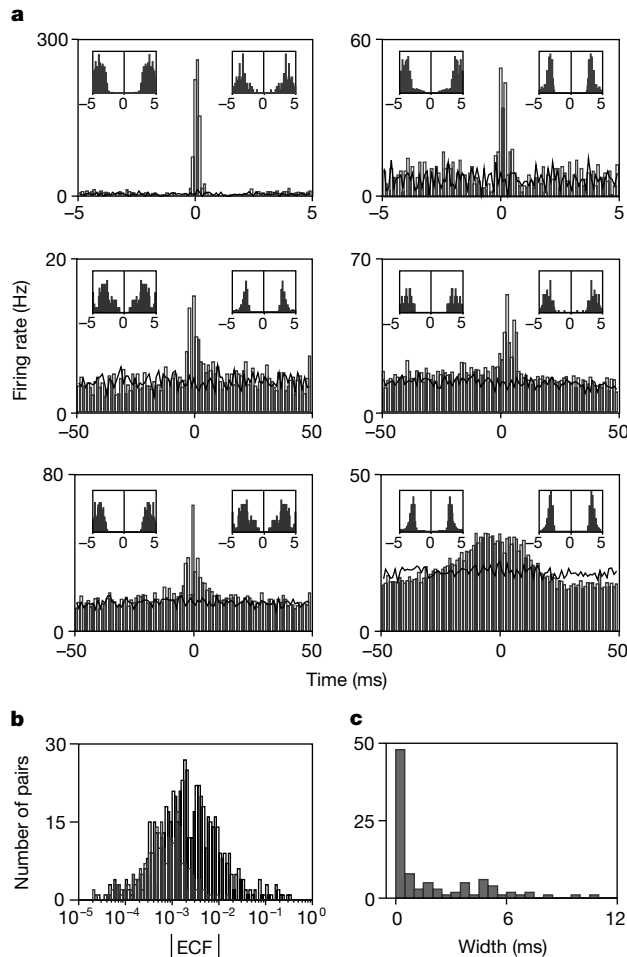


Figure 2 The degree of correlated activity and the timescale over which it occurs.

a, Cross-correlograms for pairs of ganglion cells. A cross-correlogram gives the firing rate of one cell relative to spikes generated by the other. Grey, the raw cross-correlogram; black, the shifted cross-correlogram (shift-predictor¹⁴). The latter gives the correlations produced only by the stimulus, and is generated by presenting the stimulus multiple times and cross-correlating the responses of the cells when they 'saw' the stimulus at different times. Cells are correlated if there is a peak in the raw cross-correlogram relative to the shifted. Note the expanded time scale in the top two plots. Insets, autocorrelograms for the cells in each pair; vertical axis is scaled to the height of the side peaks. **b**, Distribution of ECFs for all cell pairs. Black, pairs with positive ECFs; grey, pairs with negative ECFs. **c**, Distribution of correlation timescales, as measured by the width of the cross-correlogram, for cell pairs with ECFs above 0.5%. Widths were computed by fitting the difference between the raw and shifted cross-correlograms to a gaussian (see Supplementary Information). Characteristics of pairs of cells in **a**: top left, ON-type cells, ECF = 10.4%, $1 - \Delta I/I = 92.4\%$ (Fig. 3); top right, OFF-type cells, ECF = 1.8%, $1 - \Delta I/I = 98.4\%$; middle left, OFF-type cells, ECF = 0.8%, $1 - \Delta I/I = 97.8\%$; middle right, OFF-type cells, ECF = 4%, $1 - \Delta I/I = 91.5\%$; bottom left, OFF-type cells, ECF = 1.4%, $1 - \Delta I/I = 94.1\%$; bottom right, OFF-type cells, ECF = 0.8%, $1 - \Delta I/I = 92.2\%$.

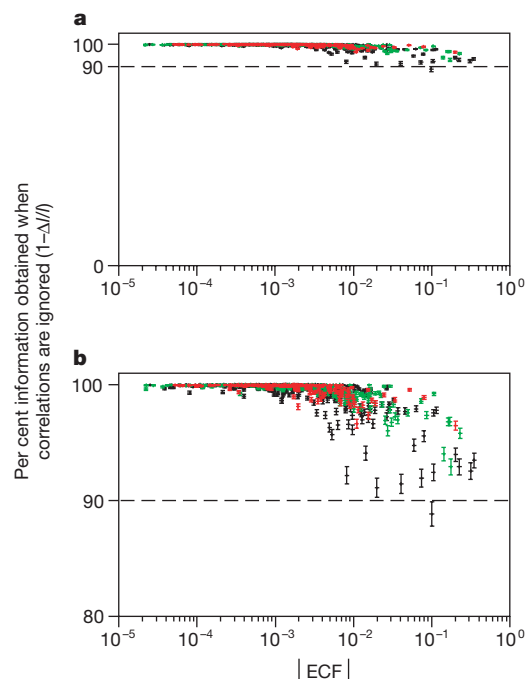


Figure 3 Per cent of information about the stimuli obtained from ganglion cell pairs when their correlated spikes were ignored, $1 - \Delta I/I$. **a**, For each pair, $1 - \Delta I/I$ was plotted against the pair's ECF. All but one pair was above the dashed line at 90%, indicating that, for nearly all pairs, > 90% of the information about the stimulus was obtained when correlations were ignored. Black, red and green correspond to Movies 1, 2, and 3, respectively. (Movie 1: 444 pairs from 5 retinas; Movie 2: 229 pairs from 3 retinas; Movie 3: 263 pairs from 3 retinas. Note that a single retina viewed multiple movies.) **b**, Expanded view.

To test the reliability of our results, we checked whether cell pairs with high ECFs were somehow anomalous. If, for example, the cells in these pairs had unusually low information rates, then the measures of ΔI might be unreliable. Similarly, if there were a large discrepancy in the information between the two cells in each pair, then the measures of ΔI might also be unreliable. For example, if one cell in the pair carried most of the information, then ΔI would be small, but reflect nothing about the importance (or lack of importance) of correlations. We tested these possibilities by measuring the information rates of both cells in all pairs and by computing the ratio of the two. No trend towards low information rates was observed for pairs of cells with high ECFs ($P > 0.3$, t -test, comparison of cell pairs above and below 0.5% ECF), and no trend towards high ratios was found ($P > 0.6$, t -test, comparison of cell pairs above and below 0.5% ECF). Finally, we measured the information rates for all cells in the dataset. These ranged from 0.15 to 4.4 bits per spike, or 0.5 to 41.2 bits per s. Compared to cat LGN, the closest

mammalian data available, the rates in our dataset were very similar in bits per spike, but a factor of two lower in bits per s (ref. 18). The lower rate is probably due to lower firing rates in our recordings.

Our results indicate that little information is lost when the correlated firing of retinal ganglion cells is ignored. This has direct bearing on the strategies that can be used for decoding ganglion cell activity. It indicates that strategies that treat the cells as independent encoders are reasonable, as they can capture more than 90% of the information that the cells carry. An example is shown in Fig. 4: we compared a decoding strategy that treated the responses of the cells as independent with one that took correlations in the responses into account, and found that both strategies performed essentially identically.

Our results also have practical implications. If ganglion cells can be treated as independent, then the activity of any given cell can be evaluated separately from other cells. If the cells cannot be treated as independent, then the activity of any given cell cannot be evaluated without taking into account other cells in the population. This greatly affects the amount of data needed to determine the information carried by a population of cells. If the cells are independent, then the amount of data needed scales linearly with the number of cells. If not, then the amount of data needed scales quadratically at best, and exponentially at worst, with the number of cells, making the problem data-limited and computationally intractable for more than a small number of cells. Our finding that ganglion cells act largely as independent encoders is thus significant from a practical, as well as scientific, point of view—it indicates that the problem of decoding population activity can be significantly simplified.

One caveat is that capturing more than 90% of the information in ganglion cell activity may ultimately not be enough. However, as our understanding of how populations of ganglion cells encode visual information is in its early stages, focusing on the part of the activity that carries most of the information seems a reasonable place to start.

The experiments in this study were performed in the light-adapted state of the retina (in daylight vision; see Methods). Further studies will be needed to determine whether correlated activity is important in the dark-adapted state, as the strategy for encoding information in the two states may not be the same. $\square$

Methods

Definition of correlated activity

'Correlated' is defined here as a lack of independence on a stimulus-by-stimulus basis. Two neurons are correlated if their conditional probabilities are not independent; that is, if $P(r_1, r_2 | s) \neq P(r_1 | s)P(r_2 | s)$ for at least one stimulus, s . Here r_1 and r_2 are the responses of cells 1 and 2, and $P(r | s)$ denotes the conditional probability of observing response r given stimulus s .

Stimuli

Three movies were used. Movies 1 and 2 contained scenes of male and female mice interacting. Second order spatial statistics were computed for these movies¹⁹; Movies 1 and 2 had $1/f^2$ and $1/f^{2.4}$ power spectra, respectively, where f is spatial frequency. These are similar to the spectra of natural scenes; stills photographed in the woods exhibit a $1/f^{1.8}$ power spectra. Movie 3 had no spatial structure, but was time-varying, with a $1/\omega$ power spectrum where ω is temporal frequency.

Mean stimulus intensity was 1,050 rod-equivalent-photons per μm^2 per s. This corresponds to 500 photoisomerizations (R^*) per rod per s, assuming that the mouse rod has an optical density at peak absorption wavelength of $0.011 \mu\text{m}^{-1}$, a length of $24 \mu\text{m}$, a diameter of $1.4 \mu\text{m}$ and a quantum efficiency of 0.67 (refs 20, 21). This stimulus intensity is in the cone regime^{22,23}, with rods $\sim 90\%$ saturated²³. The retina was also exposed briefly (< 4 min) to dim red illumination during dissection. The total equivalent photon dose during dissection was small compared to that during the stimulus: ratio of dissection to stimulus dose was 1.1% for rods and 2.7% for the longer wavelength cones (see Supplementary Information for all calculations).

Electrophysiological recording

Recordings were made with a multi-electrode array, using recording conditions described in ref. 24. To ensure that cells with overlapping receptive fields were examined, arrays with closely spaced electrodes ($25\text{--}100 \mu\text{m}$ spacing) were used. As this also increased the likelihood of recording one cell on two electrodes, the following method was used to determine whether two units were from a single cell: the receptive field of each unit was

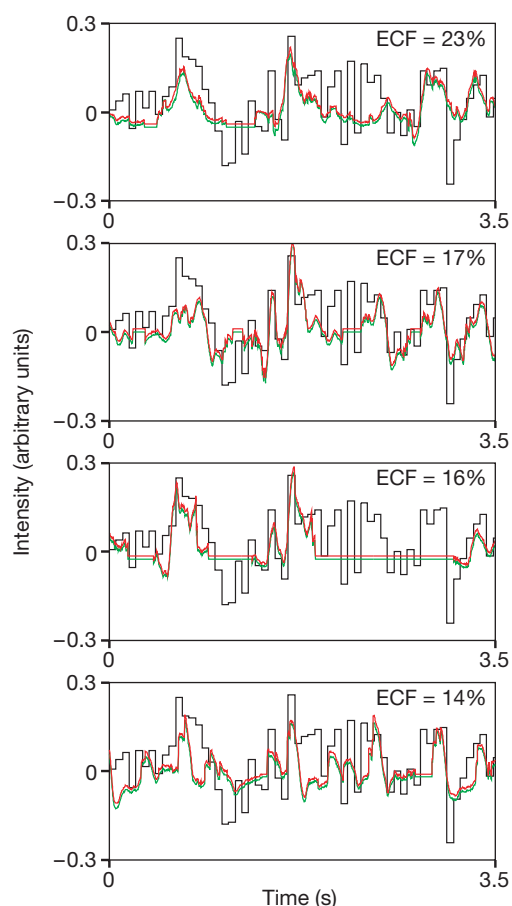


Figure 4 Decoding ganglion cell responses two ways: when the responses were treated as independent, and when correlations in the responses were taken into account. We presented the retina with a stimulus, a spatially uniform field that varied in intensity as a function of time (Movie 3), and recorded responses from pairs of cells. We then applied a widely used decoding algorithm, linear reconstruction²⁶, to reconstruct the stimulus from the responses (see Supplementary Information). We performed the reconstructions on the four pairs of cells with the highest ECFs to this stimulus. Each panel shows the reconstructions for one pair. Black trace, stimulus; red trace, reconstruction of the stimulus when the responses were treated as independent; green trace, reconstruction of the stimulus when correlations were taken into account. (The red trace is shifted up by one line width so that both traces can be seen.) In all cases, the reconstruction when the responses were treated as independent was essentially the same as that when correlations were taken into account: the root mean square error between the stimulus and the two types of reconstruction differed by $< 1\%$.

mapped using reverse correlation to a checkerboard stimulus²⁵. If the two units showed identical fields, they were considered suspect. Their spike trains were then compared to determine whether they were the same cell. Eight units showed receptive fields and spike trains identical to those of another unit (>90% identical spikes). These were discarded. Three units showed receptive fields that were identical to those of another, but failed to show identical spike trains. These were retained.

Computing the ECF

The ECF was calculated as follows: first, the raw fraction of correlated spikes for each pair was determined. This was the number of spikes that occurred within 1 ms of each other divided by the total number of spikes fired by the pair. Next, the shifted fraction was determined. This was the fraction that would have occurred by chance, taking into account correlations induced by the stimulus. It was obtained by pairing responses from the two cells when they 'saw' the stimulus at different times; that is, by pairing the response of one cell with the response of the other shifted by one movie length (the stimulus consisted of a movie repeated multiple times). We determined the shifted fraction by counting the number of spikes in the shifted pair that occurred within 1 ms of each other and dividing this by the total number of spikes for the pair. The ECF is the difference between the raw and shifted fraction.

Computing information

I and ΔI were computed as described¹⁶. A detailed description, including calculation of the error bars and demonstration that our estimate of $\Delta I/I$ is accurate, is given in Supplementary Information.

Derivation of equation (2)

Information about a stimulus can be thought of as the average number of yes/no questions it would take to identify a stimulus minus the average number of yes/no questions it would take to identify the stimulus given that responses have been observed (assuming one uses the optimal question-asking strategy; see ref. 17, chapter 5). To use the optimal strategy, one needs to know the true conditional probability distribution, $P(r_1, r_2|s)$. When we treat the cells as independent, we use an approximate conditional probability distribution, $P_{\text{IND}}(r_1, r_2|s)$. This leads to a less-than-optimal question-asking strategy, and thus a loss of information. The average number of yes/no questions needed to determine the stimulus given the responses using $P_{\text{IND}}(r_1, r_2|s)$ is $H(S|R_1, R_2) + D(P(s|r_1, r_2)||P_{\text{IND}}(s|r_1, r_2))$, where $H(S|R_1, R_2) = -\sum_{r_1, r_2} P(r_1, r_2) \log_2 P(s|r_1, r_2)$ is the conditional entropy, and $D(P(s|r_1, r_2)||P_{\text{IND}}(s|r_1, r_2)) = \sum_{r_1, r_2} P(r_1, r_2) \log_2 [P(s|r_1, r_2)/P_{\text{IND}}(s|r_1, r_2)]$ is the conditional relative entropy (see ref. 17, theorem 5.4.3, p.89). The distributions $P(s|r_1, r_2)$ and $P_{\text{IND}}(s|r_1, r_2)$ come from $P(r_1, r_2|s)$ and $P_{\text{IND}}(r_1, r_2|s)$ via Bayes' theorem. Using the standard result that the average number of yes/no questions needed to determine the stimulus when no responses have been observed is simply $H(S)$, the entropy of the stimulus¹⁷, we find that the difference, ΔI , between the information given by the true distribution and that given by the approximate distribution is $\Delta I = D(P(s|r_1, r_2)||P_{\text{IND}}(s|r_1, r_2))$, in bits. Bayes' theorem reduces this expression to equation (2).

Received 29 February; accepted 12 April 2001.

- Gray, C. M. The temporal correlation hypothesis of visual feature integration: still alive and well. *Neuron* **24**, 31–47 (1999).
- Shadlen, M. N. & Movshon, J. A. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* **24**, 67–77 (1999).
- Panzeri, S., Schultz, S. R., Treves, A. & Rolls, E. T. Correlations and the encoding of information in the nervous system. *Proc. R. Soc. Lond. B* **266**, 1001–1012 (1999).
- Nirenberg, S. & Latham, P. E. Population coding in the retina. *Curr. Opin. Neurobiol.* **8**, 488–493 (1998).
- Meister, M. Multineuronal codes in retinal signaling. *Proc. Natl Acad. Sci. USA* **93**, 609–614 (1996).
- Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
- Warland, D. K., Reinagel, P. & Meister, M. Decoding visual information from a population of retinal ganglion cells. *J. Neurophysiol.* **78**, 2336–2350 (1997).
- Rodiek, R. W. Maintained activity of cat retinal ganglion cells. *J. Neurophysiol.* **30**, 1043–1071 (1967).
- Mastroratte, D. N. Correlated firing of cat retinal ganglion cells. I. Spontaneously active inputs to X- and Y-cells. *J. Neurophysiol.* **49**, 303–324 (1983).
- Mastroratte, D. N. Correlated firing of cat retinal ganglion cells. II. Responses of X- and Y-cells to single quantal events. *J. Neurophysiol.* **49**, 325–349 (1983).
- Devries, S. H. Correlated firing in rabbit retinal ganglion cells. *J. Neurophysiol.* **81**, 908–920 (1999).
- Chichilnisky, E. J. & Baylor, D. A. Synchronized firing by ganglion cells in monkey retina. *Soc. Neurosci. Abstr.* **25**, 1042 (1999).
- Alonso, J. M., Usrey, W. M. & Reid, R. C. Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature* **383**, 815–819 (1996).
- Perkel, D. H., Gerstein, G. L. & Moore, G. P. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys. J.* **7**, 419–440 (1967).
- Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. Illinois Press, Urbana, Illinois, 1949).
- Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **80**, 197–200 (1998).
- Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley, New York, 1991).
- Reinagel, P. & Reid, R. C. Temporal coding of visual information in the thalamus. *J. Neurosci.* **20**, 5392–5400 (2000).
- Ruderman, D. L. & Bialek, W. Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.* **73**, 814–817 (1994).

- Carter-Dawson, L. D. & LaVail, M. M. Rods and cones in the mouse retina. I. Structural analysis using lights and electron microscopy. *J. Comp. Neurol.* **188**, 245–262 (1979).
- Penn, J. S. & Williams, T. P. A new microspectrophotometric method for measuring absorbance of rat photoreceptors. *Vision Res.* **24**, 1673–1676 (1984).
- Soucy, E., Wang, Y., Nirenberg, S., Nathans, J. & Meister, M. A novel signaling pathway from rod photoreceptors to ganglion cells in mammalian retina. *Neuron* **21**, 481–493 (1998).
- Dodd, R. L. in *Program in Neurosciences* 153–156 (Stanford Univ., Palo Alto, 1988).
- Nirenberg, S. & Meister, M. The light response of retinal ganglion cells is truncated by a displaced amacrine circuit. *Neuron* **18**, 637–650 (1997).
- Meister, M., Pine, J. & Baylor, D. A. Multi-neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Methods* **51**, 95–106 (1994).
- Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. Assad, A. Pouget, T. Otis, D. Buonomano and M. Goldman for critical reviews of the manuscript. We also thank M. Jack, J. Sinclair, F. Schweizer, J. Feldman and M. Meister for helpful discussion. This work was supported by grants from the Beckman Foundation and the Klingenstein Fund (S.N.).

Correspondence and requests for materials should be addressed to S.N. (e-mail: sheilan@ucla.edu).

Regulation of Ca²⁺ channel expression at the cell surface by the small G-protein kir/Gem

Pascal Béguin^{†‡}, Kazuaki Nagashima^{†‡}, Tohru Gono[§], Tadao Shibasaki[†], Kazuo Takahashi^{||}, Yasushige Kashima^{*}, Nobuaki Ozaki^{*}, Käthi Geering[†], Toshihiko Iwanaga[¶] & Susumu Seino^{*}

^{*} Department of Cellular and Molecular Medicine, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan

[†] Institut de Pharmacologie et de Toxicologie de l'Université de Lausanne, 27 rue du Bugnon, CH-1005 Lausanne, Switzerland

[§] Research Center for Pathogenic Fungi and Microbial Toxins, Chiba University, 1-8-1, Inohana, Chuo-ku, Chiba 260-8673, Japan

^{||} Department of Molecular Immunology, Core Research for Evolutional Science and Technology (CREST), Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan

[¶] Laboratory of Anatomy, Graduate School of Veterinary Medicine, Hokkaido University, Sapporo 060-0818, Japan

[‡] These authors contributed equally to this work.

Voltage-dependent calcium (Ca²⁺) channels are involved in many specialized cellular functions^{1–3}, and are controlled by intracellular signals such as heterotrimeric G-proteins⁴, protein kinases^{5,6} and calmodulin (CaM)^{7,8}. However, the direct role of small G-proteins in the regulation of Ca²⁺ channels is unclear. We report here that the GTP-bound form of kir/Gem, identified originally as a Ras-related small G-protein that binds CaM^{9–11}, inhibits high-voltage-activated Ca²⁺ channel activities by interacting directly with the β -subunit. The reduced channel activities are due to a decrease in α_1 -subunit expression at the plasma membrane. The binding of Ca²⁺/CaM to kir/Gem is required for this inhibitory effect by promoting the cytoplasmic localization of kir/Gem. Inhibition of L-type Ca²⁺ channels by kir/Gem prevents Ca²⁺-triggered exocytosis in hormone-secreting cells. We propose that the small G-protein kir/Gem, interacting with β -subunits, regulates Ca²⁺ channel expression at the cell surface.

The $\alpha_{1.2}$ - and $\alpha_{1.3}$ -subunits of L-type Ca²⁺ channels (formerly termed α_{1C} and α_{1D} , respectively¹²) are associated with auxiliary subunits (β -, $\alpha_2\delta$ - and γ -subunits) that have regulatory functions¹³.

So far four isoforms of the β -subunit (β_1 , β_2 , β_3 and β_4) have been identified¹⁴. A yeast two-hybrid screen of a complementary DNA library of mouse insulin-secreting MIN6 cells revealed a small G-protein kir (kir/Gem) that interacts directly with β_3 -subunit of L-type Ca^{2+} channels. kir/Gem, a member of the Ras-related small G-protein family, was originally cloned from mitogen-induced peripheral blood T cells⁹ and from cells transformed by *abl* tyrosine kinase oncogenes¹⁰. kir/Gem contains three important functional domains: phosphate-, guanine- and CaM-binding domains (Fig. 1a). Like other small G-proteins, kir/Gem cycles between an active GTP-bound state and an inactive GDP-bound state¹⁵. We examined the effects of guanine nucleotides on CaM binding to kir/Gem. Ca^{2+} /CaM binds to the GTP form as well as the GDP form of kir/Gem, although it has a preference for the GTP form (2.7 ± 0.1 times more GTP form than GDP form of kir/Gem synthesized in *Xenopus* oocytes was recovered on Ca^{2+} /CaM-sepharose beads, $n = 4$). In addition to the tissues previously reported⁹ (data not shown), kir/Gem messenger RNA is present at high levels in pituitary and at moderate to low levels in adrenal cells, pancreatic islets and the insulin-secreting cell lines MIN6 and RINm5F, but it is absent in brain⁹ and in rat pheochromocytoma-derived cell line PC12 (Fig. 1b).

Using a yeast two-hybrid assay, we determined the interaction of kir/Gem with the various β -subunit isoforms. We also investigated the role of the CaM-binding domain in these interactions by using two mutants: the W269G mutant, in which high-affinity interaction with CaM is disrupted¹¹, and a truncation mutant (kir/Gem ΔC)

lacking the CaM binding domain, which was originally cloned from the MIN6 cDNA library. As measured by β -galactosidase activity, all of the β -subunit isoforms examined (β_1 , β_2 and β_3) interacted with wild-type kir/Gem, the W269G mutant and the kir/Gem ΔC mutant (see Supplementary Information). The β -subunits interacted less efficiently with the wild-type kir/Gem than with the CaM-binding site mutants, indicating that CaM may interfere with the interaction between kir/Gem and the β -subunits *in vivo*.

To confirm that kir/Gem binds directly to the β -subunits, glutathione S-transferase (GST) fusion proteins of wild-type and mutant kir/Gem were immobilized on glutathione beads and incubated with homogenates of *Xenopus* oocytes expressing the β_1 -, β_2 - or β_3 -subunits (Fig. 1c). The binding of wild-type kir/Gem and the W269G mutant to the β -subunits was much more pronounced in the presence of the non-hydrolysable GTP analogue GTP- γS than of GDP- βS . Neither GST nor nucleotide-free kir/Gem bound to the β -subunits. We further confirmed the interaction between kir/Gem and β_3 -subunits *in vivo* using transfected HEK293 cells (Fig. 1d). To determine whether CaM interferes with the interaction between the β -subunits and the GTP form of kir/Gem, we included purified CaM with or without Ca^{2+} in the binding reaction. The presence of both Ca^{2+} and CaM inhibited the interaction of the β_3 -subunits with wild-type kir/Gem but not with the W269G mutant (Fig. 1e), indicating that Ca^{2+} /CaM impedes the interaction between the β -subunits and the GTP-form of kir/Gem.

To examine the functional role of kir/Gem in voltage-dependent calcium channels (VDCCs), we co-expressed β -subunits (β_1 , β_2 or

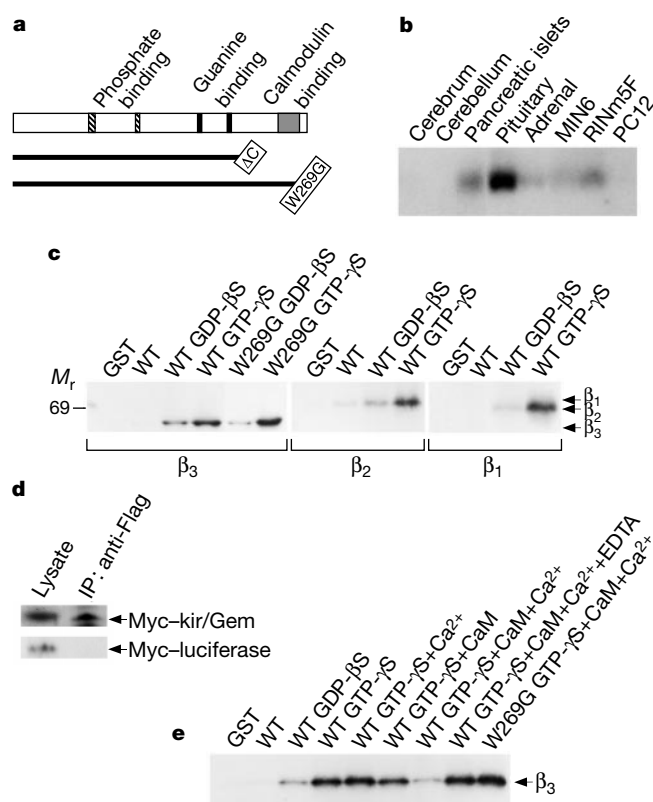


Figure 1 Characteristics of kir/Gem and its interaction with β -subunits. **a**, Functional domains of kir/Gem. ΔC , truncated form of kir/Gem lacking residues 260–295. W269G, mutant in which high-affinity CaM-binding site is disrupted. **b**, Northern blot of kir/Gem. (20 μg of total RNA from the various tissues and cell lines and 5 μg from pancreatic islets were used.) **c**, Interaction between kir/Gem and β -subunits *in vitro*. Immobilized GST–wt (wild-type) or mutant kir/Gem preloaded with GDP- βS or GTP- γS were incubated with homogenates of [^{35}S]methionine-labelled *Xenopus* oocytes expressing β_1 -, β_2 - or β_3 -subunits. **d**, Interaction of kir/Gem and the β -subunits in intact cells. HEK293 cells were

transfected with Flag-tagged β_3 -subunits and Myc-tagged kir/Gem or luciferase. Left lane, immunoblotting with anti-Myc antibody. Right lane, immunoblotting with anti-Myc antibody after immunoprecipitation (IP) with anti-Flag antibody. **e**, Inhibition of interaction between the β_3 -subunit and kir/Gem by Ca^{2+} /CaM. CaM (28 pmol) was incubated with immobilized wild-type (WT) or W269G kir/Gem (25 pmol) in the presence or absence of Ca^{2+} (2 μM) with or without EDTA (2 mM) before addition of homogenates of *Xenopus* oocytes expressing the β_3 -subunit. One of three similar experiments is shown.

β_3) with either $\alpha_1.1.2$ - or $\alpha_1.1.3$ -subunits, with or without wild-type kir/Gem in *Xenopus* oocytes (Fig. 2a–f). The co-expression of $\alpha_1.1.3$ -subunits with β_1 -, β_2 - or β_3 -subunits in the oocytes elicited Ca^{2+} channel currents (Fig. 2a–c), but not when $\alpha_1.1.3$ -subunits were expressed alone (data not shown). This confirmed that $\alpha_1.1.3$ -subunit-mediated Ca^{2+} channel currents require a β -subunit^{16,17}. Ca^{2+} channel currents were generated in about 80% of the oocytes expressing $\alpha_1.1.3$ -subunits and a β_1 -, β_2 - or β_3 -subunit. In contrast, no Ca^{2+} channel currents were recorded from the oocytes co-expressing the Ca^{2+} channel subunits and kir/Gem. Co-expression with another small G-protein, RIN¹⁸, which also binds CaM, did not abolish Ca^{2+} channel currents (Fig. 2d), indicating that the inhibitory effect is kir/Gem-specific. To determine whether a β -subunit is required for the effect of kir/Gem on Ca^{2+} channel activity, we examined the effect of kir/Gem on $\alpha_1.1.2$ -subunit-mediated Ca^{2+} channel currents, which are known to occur without a β -subunit¹⁹. These currents were not inhibited by co-expression with kir/Gem (Fig. 2e), but those generated by $\alpha_1.1.2$ -subunits and β_3 -subunits were abolished (Fig. 2f), indicating that a β -subunit is required for the inhibitory effect of kir/Gem. In addition, kir/Gem expression at non-maximal blocking doses did not alter the current–voltage (I – V) relationship and kinetics, indicating that kir/Gem does not modulate the channel gating (see Supplementary Information). We also examined whether kir/Gem had an inhibitory effect on other high-voltage-activated (HVA) Ca^{2+} channels. Overexpression of kir/Gem into BHK cells stably expressing $\alpha_2.1$ -, $\alpha_2\delta$ - and β_1 -subunits²⁰ (Fig. 2g) and into BHK cells expressing $\alpha_1.2.2$ -, $\alpha_2\delta$ - and β_1 -subunits²¹ (Fig. 2h) markedly inhibited Ca^{2+} channel currents in both groups of BHK cells. This indicated that kir/Gem inhibits both P/Q-type and N-type Ca^{2+} channels.

The β -subunits of Ca^{2+} channels have a chaperone-like role in the trafficking of α_1 -subunits to the plasma membrane^{22,23}. To investigate the effect of the interaction between the β -subunits and kir/Gem on cell surface expression of α_1 -subunits, we determined the subcellular localization of epitope-tagged $\alpha_1.1.2$ -sub-

units and kir/Gem transfected in HEK293 cells using confocal microscopy. Co-expression of $\alpha_1.1.2$ - and β_3 -subunits resulted in the expression of $\alpha_1.1.2$ -subunits at the plasma membrane (Fig. 3a), as reported elsewhere^{22,23}. By contrast, co-expression of $\alpha_1.1.2$ - and β_3 -subunits with wild-type kir/Gem, which is localized at the cytoplasmic side of the plasma membrane, blocked $\alpha_1.1.2$ -subunit cell surface expression ($\alpha_1.1.2$ -subunits aggregate intracellularly) (Fig. 3b, c). Co-expression of $\alpha_1.1.2$ - and β_3 -subunits with the W269G mutant, which is localized mainly in the nucleus, did not inhibit cell surface expression of $\alpha_1.1.2$ -subunits (Fig. 3d). These results indicate that kir/Gem, interacting with the β -subunits, interferes with the trafficking of α_1 -subunits to the plasma membrane and that Ca^{2+} /CaM contributes to this effect by promoting the cytoplasmic localization of kir/Gem. We have also found that overexpression of wild-type kir/Gem inhibits the physical association between $\alpha_1.1.2$ - and β_3 -subunits in transfected HEK293 cells (see Supplementary Information).

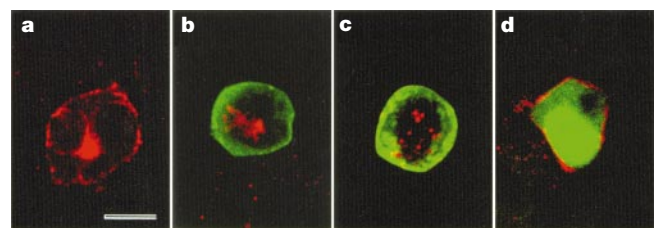


Figure 3 Subcellular localization of $\alpha_1.1.2$ -subunits, wild-type kir/Gem and the W269G mutant. HEK293 cells were transfected with HA-tagged $\alpha_1.1.2$ - and β_3 -subunits (**a**) and Myc-tagged wild-type kir/Gem (**b**, **c**), and Myc-tagged W269G mutant kir/Gem (**d**). Double immunostaining was performed using antibodies against HA (red) and Myc (green). Note that **b** and **c** are different cells. In **d**, most of the W269G mutants were located in the nucleus, as was found in a kir/Gem lacking the CaM-binding domain⁹. Scale bar, 10 μm .

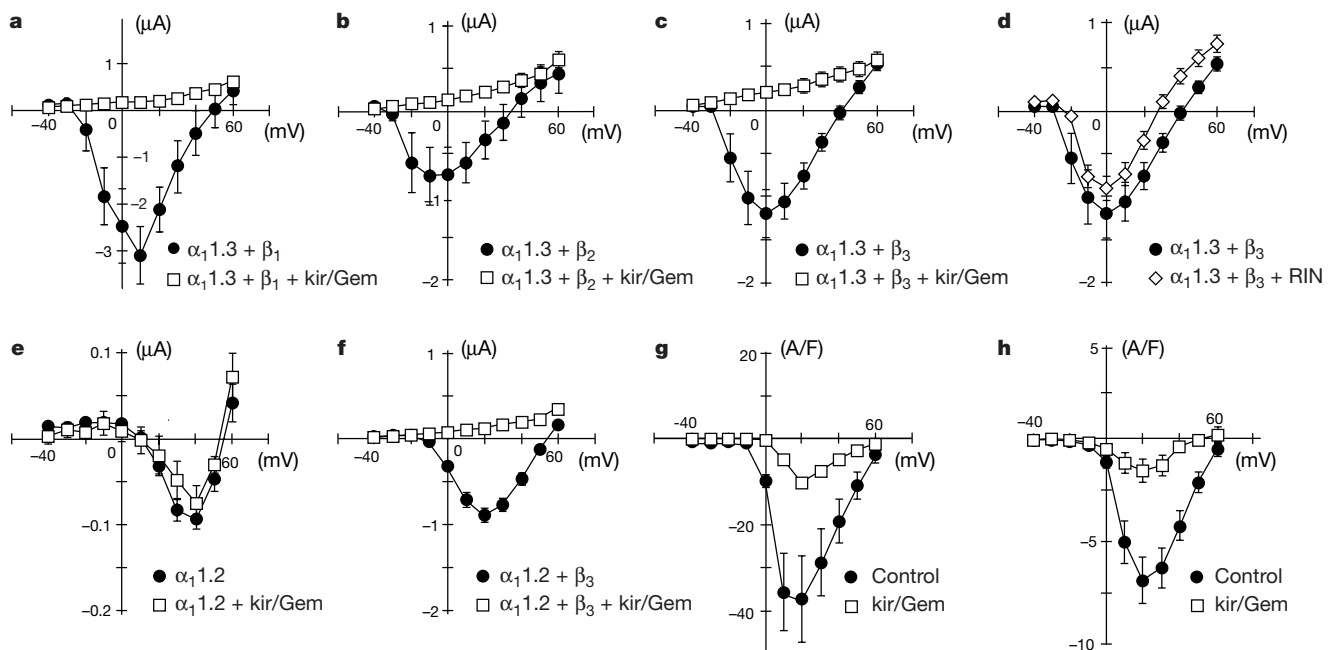


Figure 2 Effects of kir/Gem on I – V relationships of Ca^{2+} channel currents in *Xenopus* oocytes and in BHK cells. Effects of wild-type kir/Gem on I – V relationships of Ca^{2+} channels comprising $\alpha_1.1.3$ - and β_1 -subunits (**a**), $\alpha_1.1.3$ - and β_2 -subunits (**b**) or $\alpha_1.1.3$ - and β_3 -subunits (**c**). **d**, Effects of RIN on I – V relationship of Ca^{2+} channels comprising $\alpha_1.1.3$ - and β_3 -subunits. Effects of wild-type kir/Gem on I – V relationships of Ca^{2+}

channels comprising $\alpha_1.1.2$ -subunit alone (**e**) or $\alpha_1.1.2$ - and β_3 -subunits (**f**). Effects of wild-type kir/Gem on I – V relationships of Ca^{2+} channels in BHK cells expressing $\alpha_2.1$ -, $\alpha_2\delta$ -, and β_1 -subunits (**g**) or in BHK cells expressing $\alpha_1.2.2$ -, $\alpha_2\delta$ -, and β_1 -subunits (**h**). Current amplitudes are shown as mean $\pm$ s.e.m. of 10–36 different *Xenopus* oocytes or BHK cells.

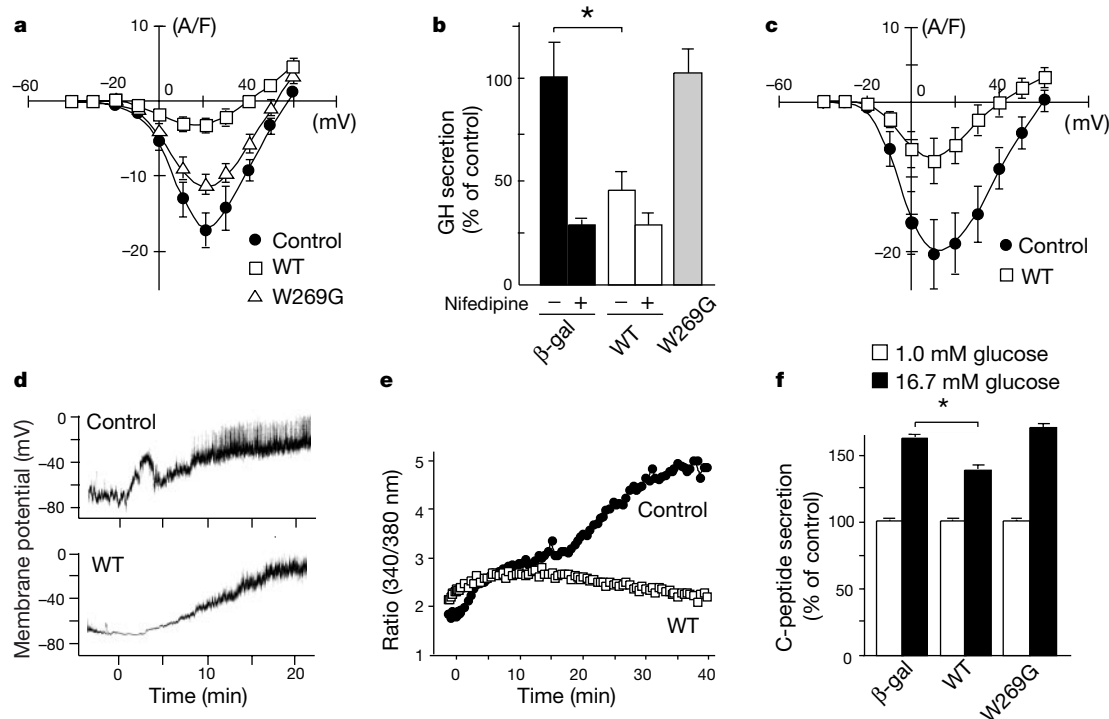


Figure 4 Effects of wild-type (WT) kir/Gem and W269G mutant in hormone-secreting cells. **a**, I-V relationships of Ca^{2+} channels in PC12 cells transfected with WT or W269G mutant kir/Gem. The maximal currents were detected at +20 mV in all cases. Control, GFP-transfected PC12 cells. **b**, Effects of kir/Gem on Ca^{2+} -triggered GH secretion in PC12 cells. PC12 cells were transfected with human growth hormone together with β -galactosidase (β -gal), WT kir/Gem or the W269G mutant. In some experiments, nifedipine (10 μM) was added before high K^+ (60 mM) stimulation. Growth hormone secretion by high K^+ stimulation in PC12 cells transfected with human growth hormone and β -gal was used as control (100%). Values are mean $\pm$ s.e.m. from 3 independent experiments ($n = 9$). Single

asterisk, $P < 0.05$. **c**, I-V relationships of Ca^{2+} channels in MIN6 cells transfected with WT kir/Gem. Control, GFP-transfected MIN6 cells. **d**, Effect of WT kir/Gem on glucose-evoked membrane potential. **e**, Effect of WT kir/Gem on glucose-induced rise in $[\text{Ca}^{2+}]_i$. In **d** and **e**, low glucose concentration (2.8 mM) was switched to high glucose concentration (25 mM) at 0 min. Representative traces are shown. Control, GFP-transfected MIN6 cells. **f**, Effects of kir/Gem on glucose-induced C-peptide secretion. C-peptide secretion at low glucose concentration (1.0 mM) was used as control (100%). High glucose- (16.7 mM) induced C-peptide secretion is expressed as % relative to control. Values are mean $\pm$ s.e.m. from 4 independent experiments ($n = 32$ for each). Single asterisk, $P < 0.001$.

To elucidate the physiological role of kir/Gem, we examined the effects of kir/Gem in PC12 and MIN6 cells in which Ca^{2+} influx through L-type Ca^{2+} channels triggers hormone secretion²⁴. Endogenous kir/Gem is absent or present at only low levels in these cells. Overexpression of wild-type kir/Gem in PC12 cells resulted in a strong reduction of the endogenous Ca^{2+} channel currents, whereas overexpression of the W269G mutant had only a slight effect (Fig. 4a). Although the W269G mutant clearly binds to the β -subunits *in vitro* (Fig. 1e), the weak inhibitory effect of the W269G mutant in intact cells is likely to be because of its nuclear localization (Fig. 3d), which prevents its interaction with the β -subunits. Similarly to the L-type VDCC blocker nifedipine, overexpression of wild-type kir/Gem in PC12 cells transfected with human growth hormone significantly inhibited Ca^{2+} -triggered hormone secretion, whereas overexpression of the W269G mutant did not (Fig. 4b). These results confirm that Ca^{2+} /CaM is required for the functional effect of kir/Gem.

To verify the involvement of kir/Gem in Ca^{2+} -triggered exocytosis, we investigated the effects of kir/Gem on electrical activity, internal calcium concentration ($[\text{Ca}^{2+}]_i$) and hormone secretion in MIN6 cells. As in PC12 cells, overexpression of wild-type kir/Gem strongly reduced endogenous Ca^{2+} channel currents in the MIN6 cells (Fig. 4c). In control MIN6 cells, high glucose concentration (25 mM) gradually depolarized the plasma membrane and generated bursts of action potentials with a concomitant rise in $[\text{Ca}^{2+}]_i$, characteristics of glucose-induced VDCC activity in pancreatic β -cells. Overexpression of wild-type kir/Gem prevented these glucose-induced bursts of action potentials and the rise in $[\text{Ca}^{2+}]_i$. Similar results were obtained by depolarizing the membrane with high K^+ stimulation (data not shown). We then examined the effect of

kir/Gem on glucose-induced insulin secretion. Pro-insulin is converted into insulin and C-peptide during the secretory process in pancreatic β -cells. Thus, we monitored glucose-induced secretion by measuring the release of human C-peptide release from MIN6 cells²⁵ that were transfected with human pro-insulin cDNA together with β -galactosidase, wild-type kir/Gem or the W269G mutant. Overexpression of wild-type kir/Gem in MIN6 cells significantly suppressed glucose-induced C-peptide secretion, whereas overexpression of the W269G mutant did not (Fig. 4f). These results indicate that kir/Gem may have a physiological role in Ca^{2+} -triggered exocytosis by controlling the activity of L-type VDCCs.

On the basis of these data, we propose a model for the regulation of HVA Ca^{2+} channel expression at the cell surface by the small G-protein kir/Gem (Fig. 5). Upon activation of CaM by intracellular calcium, Ca^{2+} /CaM binds to kir/Gem, and promotes kir/Gem localization to the cytoplasm. Because the β -subunits bind preferentially to the active GTP form of kir/Gem (GTP-kir/Gem) (Fig. 1c) and Ca^{2+} /CaM inhibits this interaction (Fig. 1e), Ca^{2+} /CaM should dissociate from kir/Gem, allowing GTP-kir/Gem to interact with the β -subunit. In addition to Ca^{2+} /CaM, other factors such as guanine nucleotide exchange factors (GEFs), which favour the GTP-form of small G proteins¹⁵, may also be critical in the interaction of kir/Gem with the β -subunits. The I-II loop of the α_1 -subunit contains an endoplasmic reticulum retention signal that severely restricts cell surface expression²⁶. The β -subunit reverses the inhibition imposed by the retention signal²⁶. As we have found that overexpression of wild-type kir/Gem inhibits the physical association between α_1 1.2- and β_3 -subunits (see Supplementary Information), it is likely that kir/Gem, interacting with the β -subunits, antagonizes the reversal effect of the β -subunits, thereby

Plasma membrane

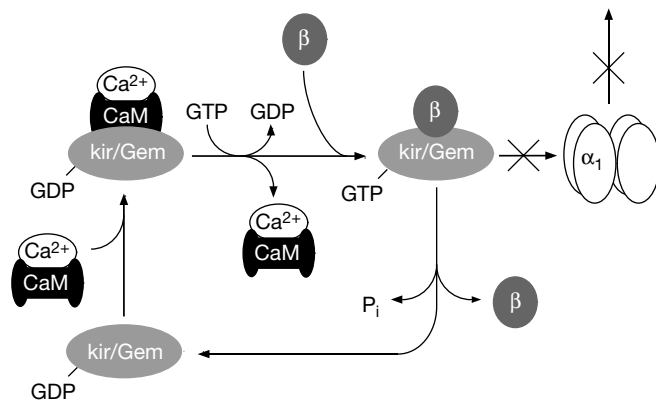


Figure 5 Model of regulation of Ca^{2+} channel expression at the cell surface by kir/Gem. Upon activation of CaM by intracellular calcium, Ca^{2+} /CaM binds to the GDP-kir/Gem form (inactive form) and localizes kir/Gem in the cytoplasm. kir/Gem then dissociates from CaM, and is converted into the active GTP form (GTP-kir/Gem) by a GEF for kir/Gem which has yet to be identified. kir/Gem then binds to a β -subunit. Binding of kir/Gem to the β -subunit interferes with the interaction between the β -subunit and the α_1 -subunit, blocking trafficking of the α_1 -subunit to the plasma membrane, and reduces the number of functional Ca^{2+} channels at the cell surface. Hydrolysis of GTP probably causes kir/Gem to dissociate from the β -subunit.

inhibiting the traffic of the α_1 -subunits to the plasma membrane.

Mechanisms of both positive and negative short-term regulation of HVA Ca^{2+} channels by Ca^{2+} have been proposed^{6–8}. Here, we describe long-term regulation of HVA Ca^{2+} channels by the calmodulin-binding small G-protein kir/Gem involving cell surface expression. As overexpression of kir/Gem in the hormone-secreting cell lines PC12 and MIN6 significantly inhibits Ca^{2+} -triggered exocytosis, kir/Gem may participate in the regulation of hormone release. Alternatively, kir/Gem could have a protective effect against the overload of Ca^{2+} (Ca^{2+} toxicity) that impairs a variety of cellular functions. kir/Gem might also inhibit cell proliferation by reducing Ca^{2+} -mediated cell growth^{9,27}. □

Methods

Molecular biology

The yeast two-hybrid screen of a mouse MIN6 cDNA library using the β_3 -subunit (amino-acid residues 50–484) as bait was performed as described²⁵. Site-directed mutagenesis on kir/Gem was done using the polymerase chain reaction (PCR)-based method. Total RNA was blotted under standard stringent hybridization and washing conditions using a full-length kir/Gem cDNA as probe. *In vitro* synthesized cRNAs (7 ng) encoding the rat $\alpha_1.3$ -subunit or the rabbit $\alpha_1.2$ -subunit were co-injected with 3 ng of different rat β -subunit cRNAs with or without kir/Gem cRNA (0.3 ng) in stage V–VI *Xenopus* oocytes. Cell culture and DNA transfection were carried out as described^{25,28}.

Biochemistry

GST–kir/Gem fusion proteins were obtained as described¹¹. Wild-type or mutant GST–kir/Gem proteins (amino-acid residues 20–295) were loaded with GDP- β S or GTP- γ S and immobilized on glutathione beads¹¹ (Sigma). Immobilized GST–kir/Gem was incubated with [³⁵S]methionine-labelled homogenates of *Xenopus* oocytes²⁸ in a solution of 50 mM Tris-HCl buffer (pH 7.5), 0.1 mM dithiothreitol, 1 mM MgCl_2 , 100 mM NaCl and 0.05% Tween 20 for 3 h at 4 °C. The bound proteins were analysed by SDS–PAGE and revealed by fluorography. Bovine brain calmodulin (Calbiochem) was added during the immobilization of GST–kir/Gem on glutathione beads. The GDP or GTP bound form of kir/Gem was incubated with Ca^{2+} /CaM-sepharose, and then immunoblotted using rabbit polyclonal antibody raised against GST–kir/Gem. HEK293 cells were transfected with plasmid vectors for Myc-tagged kir/Gem and Flag-tagged β_3 -subunit. The cellular lysates were subjected to co-immunoprecipitation using anti-Flag antibody (Sigma). The proteins were analysed by immunoblotting with anti-Myc antibody (Sant Cruz Biotechnology).

Electrophysiology

A conventional two-electrode voltage clamp method was used in *Xenopus* oocytes. Bath solution for *Xenopus* oocytes was described elsewhere⁸. Recordings also were made in

PC12, MIN6 and BHK cells with the whole-cell patch clamp method using bath solution with 40 mM Ba^{2+} (refs 17 and 29). The currents (I_{Ba}) were normalized by dividing by the membrane capacitance for each cell²⁹. The holding potential was –60 mV, and test pulses of 400 ms at potentials between –40 and +60 mV in steps of 10 mV were applied every 4 s.

Immunocytochemistry

HEK293 cells were transfected with haemagglutinin (HA)-tagged $\alpha_1.2$ -subunit, β_3 -subunit and Myc-tagged wild-type kir/Gem or Myc-tagged W269G mutant, and in some cases, transfected with HA-tagged $\alpha_1.2$ - and β_3 -subunits. The cultured cells were fixed with the paraformaldehyde method. After pretreatment with Triton X-100, they were incubated with rabbit anti-HA polyclonal antisera (Zymed), then with Cy3-labelled donkey anti-rabbit IgG (Jackson ImmunoResearch Laboratories), and then with mouse anti-Myc antibody, followed by fluorescein isothiocyanate (FITC)-labelled donkey anti-guinea pig IgG (Jackson ImmunoResearch Laboratories). The sections were observed under a confocal laser scanning microscope (Fluoview, Olympus).

Measurements of growth hormone, C-peptide and $[\text{Ca}^{2+}]_i$

Human growth hormone secretion in PC12 cells and C-peptide secretion in MIN6 cells were assayed as described²⁵. $[\text{Ca}^{2+}]_i$ in MIN6 cells were measured by a dual-excitation wavelength method (340/380 nm)²⁹.

Received 2 January; accepted 22 March 2001.

- Wheeler, D. B., Randall, A. & Tsien, R. W. Roles of N-type and Q-type Ca^{2+} channels in supporting hippocampal synaptic transmission. *Science* **264**, 107–111 (1994).
- Artalejo, C. R., Adams, M. E. & Fox, A. P. Three types of Ca^{2+} channels trigger secretion with different efficacies in chromaffin cells. *Nature* **367**, 72–76 (1994).
- Hardingham, G. E., Chawla, S., Johnson, C. M. & Bading, H. Distinct functions of nuclear and cytoplasmic calcium in the control of gene expression. *Nature* **385**, 260–265 (1997).
- Dolphins, A. Mechanism of modulation of voltage-dependent calcium channels by G proteins. *J. Physiol.* **506**, 3–11 (1998).
- Dolphins, A. Facilitation of Ca^{2+} current in excitable cells. *Trends Neurosci.* **19**, 35–43 (1996).
- Dzhura, I., Wu, Y., Colbran, R. J., Balser, J. R. & Anderson, M. E. Calmodulin kinase determines calcium-dependent facilitation of L-type calcium channels. *Nature Cell Biol.* **2**, 173–177 (2000).
- Lee, A. et al. Ca^{2+} /calmodulin binds to and modulates P/Q-type calcium channels. *Nature* **399**, 155–159 (1999).
- Zühlke, R. D., Pitt, G. S., Deisseroth, K., Tsien, R. W. & Reuter, H. Calmodulin supports both inactivation and facilitation of L-type calcium channels. *Nature* **399**, 159–162 (1999).
- Maguire, J. et al. Gem: An induced, immediate early protein belonging to the Ras family. *Science* **265**, 241–244 (1994).
- Cohen, L. et al. Transcriptional activation of ras-like gene (*kir*) by oncogenic tyrosine kinases. *Proc. Natl Acad. Sci. USA* **91**, 12448–12452 (1994).
- Fischer, R., Wei, Y., Anagli, J. & Berchtold, M. W. Calmodulin binds to and inhibits GTP binding of the Ras-like GTPase Kir/Gem. *J. Biol. Chem.* **271**, 25067–25070 (1996).
- Ertel, E. A. et al. Nomenclature of voltage-gated calcium channels. *Neuron* **25**, 533–535 (2000).
- Walker, D. & De Waard, M. Subunit interaction sites in voltage-dependent Ca^{2+} channels: role in channel function. *Trends Neurosci.* **21**, 148–154 (1998).
- Isom, L. L., De Jongh, K. S. & Catterall, W. A. Auxiliary subunits of voltage-gated ion channels. *Neuron* **12**, 1183–1194 (1994).
- Cherfils, J. & Chardin, P. GEFs: structural basis for their activation of small GTP-binding protein. *Trends Biochem. Sci.* **24**, 306–311 (1999).
- Williams, M. E. et al. Structure and functional expression of $\alpha_1.1$, α_2 , and β -subunits of a novel human neuronal calcium channel subtype. *Neuron* **8**, 71–84 (1992).
- Ihara, Y. et al. Molecular diversity and functional characterization of voltage-dependent calcium channels (CACN4) expressed in pancreatic β -cells. *Mol. Endocrinol.* **9**, 121–130 (1995).
- Wes, P. D., Yu, M. & Montell, C. RIC, a calmodulin-binding Ras-like GTPase. *EMBO J.* **15**, 5839–5848 (1996).
- Mikami, A. et al. Primary structure and functional expression of cardiac dihydropyridine-sensitive calcium channel. *Nature* **340**, 230–233 (1989).
- Niidome, T. et al. Stable expression of the neuronal BI (class A) calcium channel in baby hamster kidney cells. *Biochem. Biophys. Res. Commun.* **203**, 1821–1827 (1994).
- Wakamori, M. et al. Functional characterization of ion permeation pathway in the N-type Ca^{2+} channel. *J. Neurophysiol.* **79**, 622–634 (1998).
- Chien, A. J. et al. Roles of a membrane-localized β -subunit in the formation and targeting of functional L-type Ca^{2+} channels. *J. Biol. Chem.* **270**, 30036–30044 (1995).
- Brice, N. L. et al. Importance of the different β -subunits in the membrane expression of the α_1A and α_2 calcium channel subunits: studies using a depolarization-sensitive α_1A antibody. *Eur. J. Neurosci.* **9**, 749–759 (1997).
- Plummer, M. R., Logothetis, D. E. & Hess, P. Elementary properties and pharmacological sensitivities of calcium channels in mammalian peripheral neurons subunit. *Neuron* **2**, 1453–1463 (1989).
- Ozaki, N. et al. cAMP-GEFII is a direct target of cAMP in regulated exocytosis. *Nature Cell Biol.* **2**, 805–811 (2000).
- Bichet, D. et al. The I-II loop of the Ca^{2+} channel α_1 subunit contains an endoplasmic reticulum retention signal antagonized by the β subunit. *Neuron* **25**, 177–190 (2000).
- Berridge, M. J., Bootman, M. D. & Lipp, P. Calcium—a life and death signal. *Nature* **395**, 645–648 (1998).
- Béguin, P., Nagashima, K., Nishimura, M., Gono, T. & Seino, S. PKA-mediated phosphorylation of the human K_{ATP} channel: separate roles of Kir6.2 and SUR1 subunit phosphorylation. *EMBO J.* **18**, 4722–4732 (1999).
- Kawaki, J. et al. Unresponsiveness to glibenclamide during chronic treatment induced by reduction of ATP-sensitive K^{+} channel activity. *Diabetes* **48**, 2001–2006 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank T. Tanabe and Y. Mori for rabbit α_1 1.2 cDNA and BHK cell lines expressing Ca^{2+} channels, respectively; and Y. Takai for reading the manuscript. This work was supported by a Grant-in-Aid for Creative Basic Research from the Ministry of Education, Science, Sports and Culture Japan; by Scientific Research Grants from the Ministry of Health and Welfare, Japan; and by grants from Novo Nordisk Pharma, Japan; Yamanouchi Foundation for Research on Metabolic Disorders, Japan; and Swiss National Fund for Scientific Research. P.B. and K.N. are supported by Japan Society for the Promotion of Science Fellowships for Foreign Researchers and for Young Scientists, respectively.

Correspondence and requests for materials should be addressed to S.S. (e-mail: seino@med.m.chiba-u.ac.jp).

KANADI regulates organ polarity in *Arabidopsis*

Randall A. Kerstetter, Krista Bollman, R. Alexandra Taylor, Kirsten Bombliès & R. Scott Poethig

University of Pennsylvania, Department of Biology, Plant Science Institute, Philadelphia, Pennsylvania 19104-6018, USA

Leaves and floral organs are polarized along their adaxial–abaxial (dorsal–ventral) axis. In *Arabidopsis*, this difference is particularly obvious in the first two rosette leaves, which possess trichomes (leaf hairs) on their adaxial surface but not their abaxial surface^{1–3}. Mutant alleles of *KANADI* (*KAN*) were identified in a screen for mutants that produce abaxial trichomes on these first two leaves. *kan* mutations were originally identified as enhancers of the mutant floral phenotype of *crabs claw* (*crc*), a gene that specifies abaxial identity in carpels^{4,5}. Here we show that *KAN* is required for abaxial identity in both leaves and carpels, and encodes a nuclear-localized protein in the GARP family of putative transcription factors. The expression pattern of *KAN* messenger RNA and the effect of ectopically expressing *KAN* under the regulation of the cauliflower mosaic virus (CAMV) 35S promoter indicate that *KAN* may also specify peripheral identity in the developing embryo.

The first two leaves of *kan* mutant seedlings emerge in an upright position and remain cupped throughout their development (Fig. 1a). Leaves at higher nodes also emerge as in-rolled primordia, but are almost completely flat at maturity (Fig. 1b). In contrast, wild-type leaves of the ecotype Columbia (Col) always curl downwards (Fig. 1b). This difference in leaf expansion is not associated with a change in leaf size (data not shown) but is reflected in the internal anatomy of mutant and wild-type leaves. In paradermal section, cells in the adaxial two mesophyll layers of wild-type leaves appear large, round and densely packed, whereas cells in the abaxial two mesophyll layers are irregularly shaped and are separated by large air spaces (Fig. 1c). In cross-sections of the leaf blade this is evident as a steep adaxial–abaxial gradient in the cross-sectional area and density of mesophyll cells (Fig. 1d). In *kan*, abaxial mesophyll cells are more regularly shaped (Fig. 1c) and the adaxial–abaxial gradient in cell size and cell density is less pronounced (Fig. 1d) than in wild-type leaves. This decrease in abaxial–adaxial polarity arises from a decrease in the cross-sectional area of adaxial mesophyll cells and an increase in the cross-sectional area and the number of cells in abaxial cell layers.

kan has a similar effect on trichome density: decreasing trichome number on the adaxial surface of rosette leaves while increasing abaxial trichome number (Fig. 1e). Its effect on abaxial trichome production is particularly obvious in the first few rosette leaves because these leaves do not normally produce abaxial trichomes^{2,3}. Wild-type plants first produce abaxial trichomes on leaves 5 or 6,

whereas plants homozygous for *kan* produce abaxial trichomes starting with leaf 1 or 2. Plants heterozygous for mutant alleles of *kan*, including the null alleles *kan-12* and *kan-14*, produce abaxial trichomes starting with leaf 3. This semi-dominant effect implies that *kan* is haplo-insufficient and suggests that leaf polarity may be exquisitely sensitive to the expression level of this gene. *kan* has little or no effect on adaxial–abaxial differences in epidermal cell size and shape or stomatal density, nor does it affect trichome production on inflorescence leaves, which lack trichomes on their adaxial but not their abaxial surface. Thus, *KAN* is required for some, but not all, aspects of adaxial–abaxial polarity.

In the Landsberg *erecta* (Ler) and Wassilewskije (Ws) genetic backgrounds, *kan* has floral defects similar to those described for *kan-2 crc* double mutants⁵. These include proliferation of replum tissue, production of ectopic ovules and formation of projections of carpel, style, or stigmatic tissue from the base of the pistil and the replum (Fig. 1f). This floral phenotype was apparent in many, but not all, mutant flowers in these genetic backgrounds and was suppressed in Col. The formation of ovules on the external surface of the carpel and the proliferation of the replum are consistent with

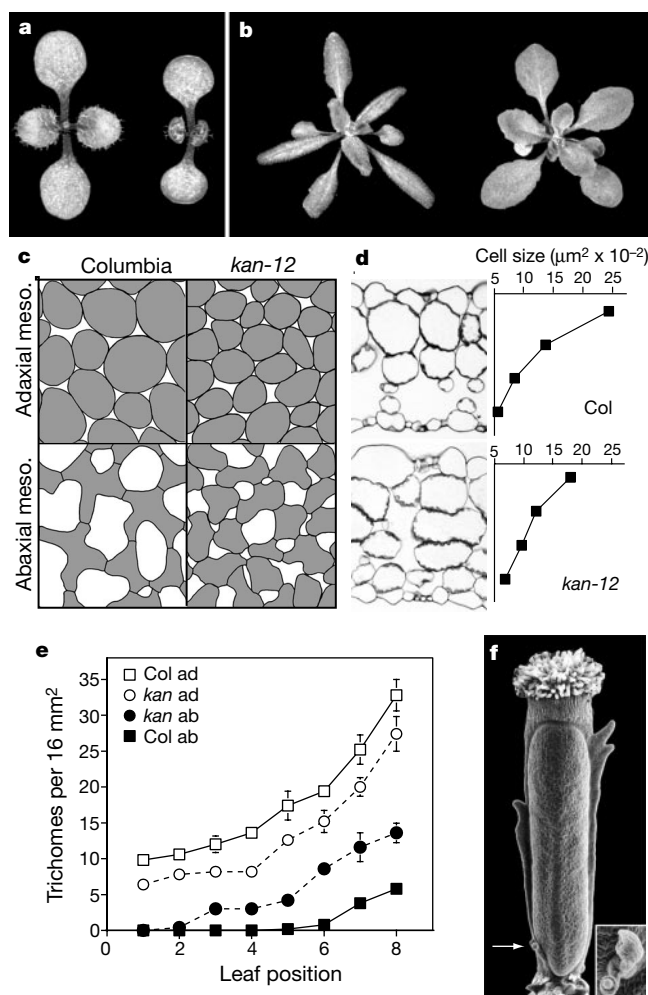


Figure 1 Effects of *kanadi* on organ polarity. **a, b**, Wild-type Col and *kan-12* plants 7 days (**a**) and 3 weeks (**b**) after planting. **c**, Camera lucida drawings of paradermal views of the adaxial and abaxial mesophyll layers of leaf 1 in Col and *kan-12*. **d**, Transverse sections through Col (top) and *kan-12* (bottom) leaves, and the cross-sectional area of cells in each of the four mesophyll layers; standard error bars are obscured by the plot symbols. **e**, Trichome density on the abaxial (ab) and adaxial (ad) surfaces of rosette leaves of Col and *kan-11* plants. **f**, Scanning electron micrograph of a pistil from *kan-12* in the Ler genetic background. Inset, magnified view of ectopic ovules (arrow).

the partial transformation of the abaxial (external) side of the pistil into adaxial (internal) tissue.

KAN was mapped close to the simple sequence length polymorphisms marker nga106 on chromosome 5, and was cloned using the T-DNA-tagged allele, *kan-13*. A 9 kilobase (kb) *SalI* fragment adjacent to the left border of the T-DNA was cloned by plasmid rescue⁶, and used to identify complementary DNAs⁷ corresponding to the gene disrupted by the T-DNA insertion. Five other *kan* alleles from different mutant screens were also found to contain mutations in this gene (Fig. 2a). KAN consists of 6 exons, and encodes a protein with a 56 amino-acid domain (residues 220–275) that shares between 30% and 98% amino-acid identity with more than 50 predicted genes in the *Arabidopsis* genome, including several putative transcription factors. This conserved domain has recently been named the GARP domain⁸ because it is present in GOLDEN2 in maize⁹, *Arabidopsis* response-regulator (ARR) proteins¹⁰ and the Psr1 protein from *Chlamydomonas*¹¹. The GARP domains of ARR1 and ARR2 have recently been shown to function in sequence-specific DNA binding, indicating that these proteins may operate as transcriptional activators¹². Although many of the proteins containing GARP domains have been classified as two-component response regulators (type B)^{10,12,13} and pseudo-response regulators^{14,15}, KAN does not contain the amino-terminal receiver domain conserved in this class of proteins. However, the predicted KAN peptide is serine/threonine-rich (21.1% by residue number) and contains as many as 46 potential phosphorylation sites¹⁶, which may serve an analogous function to the receiver domain in the ARR proteins.

Of the many GARP proteins in the *Arabidopsis* genome, three are nearly identical to KAN within the GARP domain (Fig. 2b). Although the overall identity between these proteins is low (29–38%), they contain several stretches of amino-acid identity that are not found in other family members (not shown). We have named these KANADI-like predicted genes *KNDL*s. The sequence similarity found in this subset of GARP genes may indicate that they have overlapping or partially redundant functions.

The hypothesis that KAN functions as a transcription factor is supported by its subcellular localization. Translational fusions between the KAN cDNA and either β -glucuronidase (GUS) or modified green fluorescent protein (mGFP5) were introduced into onion epidermal cells by biolistic transformation¹⁷. In both cases, KAN directed reporter gene expression to the nucleus (Fig. 2c, e). GUS and GFP expression were cytoplasmic in cells transformed with constructs lacking KAN (Fig. 2d, f).

RNA *in situ* hybridization reveals that KAN is weakly expressed throughout the early globular embryo (Fig. 3a) and soon thereafter becomes restricted to peripheral cells in a basal region of the embryo just above the hypophysis (Fig. 3b). In heart-stage embryos, expression was observed both in the periphery of the presumptive hypocotyl and on the abaxial side of cotyledon primordia (Fig. 3c). During vegetative growth, KAN mRNA was observed on the abaxial side of very young leaf primordia, but not in older leaves. It was also present transiently on the abaxial side of initiating floral-organ primordia (Fig. 3f, h) and in nectaries (not shown). A more complex pattern of expression was observed in carpels. Expression was first observed on the abaxial side of carpel primordia (Fig. 3f), and then in a localized region on the abaxial margin that gives rise to the septum (Fig. 3h). Later, KAN was expressed in the tissue that gives rise to ovules (Fig. 3g, h). The observed expression pattern is consistent with a function for KAN in either promoting abaxial or inhibiting adaxial fate during lateral organ formation, and suggests a connection between the abaxial side of cotyledon primordia and peripheral tissue in the embryo. This abaxial expression pattern also suggests that the effect of *kan* on adaxial identity (Fig. 1c–e) is indirect, and may be a consequence of the loss of abaxial identity in mutant plants.

To test the hypothesis that KAN is involved in the regulation of organ polarity, a construct containing the 'constitutive' CAMV 35S promoter fused to the KAN cDNA was introduced into plants by *Agrobacterium*-mediated transformation¹⁸. Primary, kanamycin-resistant transformants showed a marked and usually terminal seedling phenotype. In one experiment with 217 transgenic seedlings, 92.2%

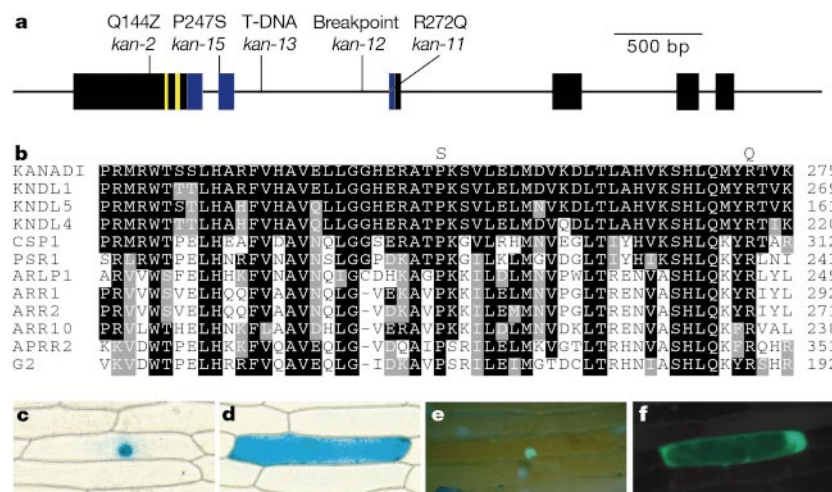


Figure 2 Molecular characterization of the *KAN* gene. **a**, *KAN* consists of 6 exons (rectangles) with a single ORF predicted to encode a protein of 403 amino acids (relative molecular mass of 45,800 (M_r 45.8K)) with a domain (blue) conserved in a large number of plant genes. Two short clusters of histidine residues are shown in yellow. *kan-2* is a nonsense mutation in exon 1; *kan-11* and *kan-15* are missense mutations in the conserved domain; *kan-12* and *kan-14* are large deletions that remove 12.1 kb (promoter and first two exons) and 13.6 kb (entire *KAN* ORF and two neighbouring genes), respectively; and *kan-13* is a T-DNA insertion allele. **b**, Fifty-six amino-acid domain in KAN conserved in at least 50 *Arabidopsis* predicted proteins, three of which (*KNDL1*, *KNDL5* and *KNDL4* (GenBank accession numbers AAG60180, BAB10501 and Z97344,

respectively)) are nearly identical to KAN in this region. Also shown are similar domains from the *Mesembryanthemum crystallinum* protein CSP1 (AAF32350 (ref. 25)), PSR1 from *Chlamydomonas* (AAD55941 (ref. 11)); G2 from maize (AG32325 (ref. 9)) and the two-component response regulators ARR1, ARR2 and ARR10 (accession numbers BAA74528, BAA74527 (refs 10, 12) and CAA06432 (ref. 13), respectively); and the pseudo-response regulators ARLP1 and APRR2 (CAA06431 (ref. 14) and BAA94548 (ref. 15)) from *Arabidopsis*. The missense mutations in *kan-15* (S) and *kan-11* (Q) are shown. **c–f**, Onion epidermal cells transformed with the *KAN* coding region fused in-frame to the N terminus of *GUS* (**c**), with 35S–*GUS* (**d**), with the *KAN* coding region fused in-frame to the N terminus of *mGFP5* (**e**), and with 35S–*mGFP5* (**f**).

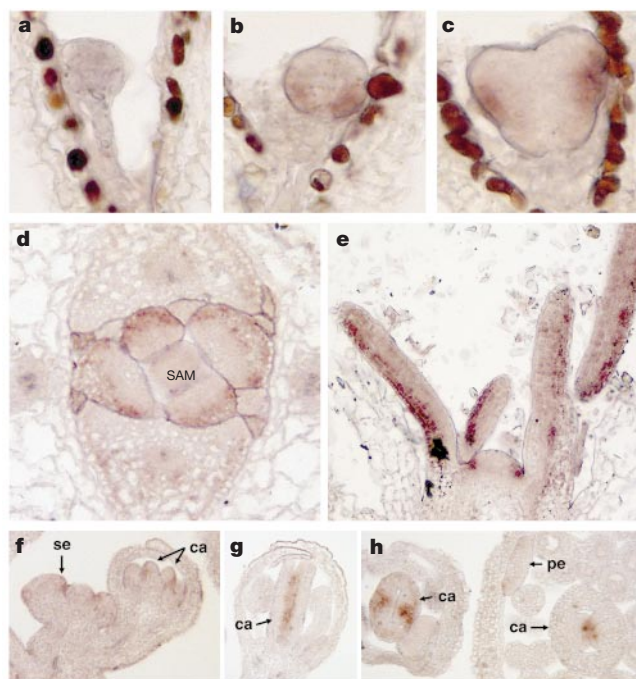


Figure 3 *In situ* location of *KAN* mRNA. **a**, Early-globular-stage embryo; **b**, late-globular-stage-embryo; **c**, early-heart-stage embryo; **d**, transverse and **e**, longitudinal sections of 7-day-old seedlings; **f**, inflorescence apex with stage-3 and stage-7 flowers; **g**, longitudinal section through a stage-9 flower; **h**, transverse section through stage-8 and stage-9 flowers. No hybridization was detected either with the sense control probe or with the anti-sense probe on tissues of plants homozygous for a deletion allele, *kan-12*. SAM, shoot apical meristem; se, sepal; ca, carpel; pe, petal.

had elongated and pointed cotyledons (Fig. 4b) most of which (89%) produced no subsequent leaves. These seedlings had rectangular, rather than irregular, cotyledon epidermal cells (Fig. 4a, b, inset). Several of the remaining transformants produced normal plants or plants with a *kan* phenotype, but most had a variety of other abnormalities including narrow or cylindrical leaves (Fig. 4e) with few or no trichomes and no vasculature, single cotyledons, or fused cotyledons. The internal tissue of transgenic cotyledons resembled abaxial spongy mesophyll in comprising small cells with a large amount of intercellular air space (Fig. 4d). Transgenic cotyledons also had a reduced amount of vascular tissue and abnormally thick margins, consistent with the absence of well-defined adaxial and abaxial domains. In addition, transgenic seedlings frequently lacked both a shoot apical meristem (SAM) and vascular tissue in the hypocotyl (Fig. 4g). Although affected hypocotyls had a normal number of cell layers, cells in the centre of the hypocotyl were unusually large and did not have the characteristic anatomy of phloem or xylem cells (Fig. 4j). The effect of *35S-KAN* on vascular differentiation was limited to the cotyledons and hypocotyl, as the root anatomy of transgenic seedlings was essentially normal (Fig. 4k). In contrast to the dramatic phenotype of *35S-KAN* seedlings, the anatomy of *kan-12* SAMs and hypocotyls was indistinguishable from wild type (Fig. 4i).

The loss-of-function phenotype of *KAN*, its expression pattern, and the phenotype of *35S-KAN* seedlings all support the conclusion that *KAN* promotes abaxial identity. In addition, the expression pattern of *KAN* and the phenotype of *35S-KAN* seedlings raise the possibility that *KAN* may also be involved in regulating peripheral identity in the embryo. The SAM and vascular tissue arise from a central domain of the embryo, defined molecularly by the expression pattern of *PINHEAD*¹⁹. Ectopic expression of a gene that

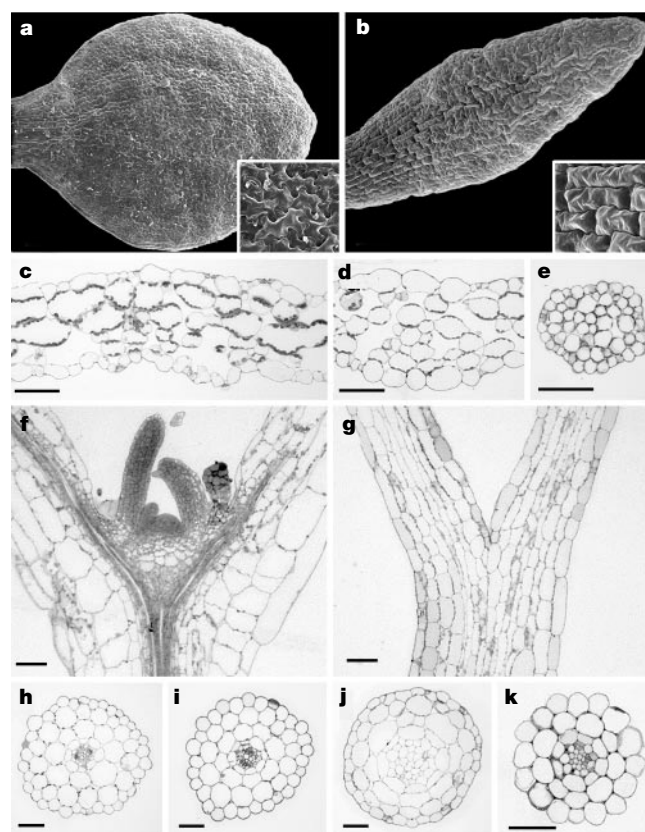


Figure 4 Seedling phenotypes of plants resulting from ectopic expression of *KAN*. **a**, **b**, Scanning electron micrographs of the abaxial sides of cotyledons from Col (**a**) and *35S-KAN* (**b**) seedlings. Insets, magnifications of the blade region to highlight epidermal cell-shape differences. **c**–**e**, Sections through expanded cotyledons of Col (**c**) and *35S-KAN* (**d**), and through a rare radial leaf (**e**) from a *35S-KAN* seedling. **f**, **g**, Longitudinal sections through 8-day-old Col (**f**) and *35S-KAN* (**g**) seedlings. **h**–**k**, Transverse sections through Col (**h**), *kan-12* (**i**) and *35S-KAN* (**j**) hypocotyls, and a *35S-KAN* (**k**) root. Scale bar, 50 μ m.

promotes peripheral cell identity would therefore be expected to affect the differentiation of both of these structures. We propose that *35S-KAN* seedlings lack SAM and vascular tissue in the hypocotyl because *35S-KAN* peripheralizes the embryo, eliminating the cell type from which both the SAM and the vascular tissue of the hypocotyl arise. Together with previous studies^{19,20}, these results suggest that the specification of adaxial–abaxial polarity in lateral organs is intimately linked to the specification of central–peripheral identity in the shoot. Given that abaxial tissue in lateral organs is spatially contiguous with peripheral tissue in the shoot axis, and that adaxial tissue is contiguous with central tissue, we believe that the radial polarity of the shoot axis and the adaxial–abaxial polarity of lateral organs may be specified by a single patterning mechanism. Future studies should provide a critical test of this hypothesis and reveal the role of *KAN* and related genes in this process. □

Methods

Plant materials

We isolated the mutant alleles *kan-11* to *kan-15* in this study. Complementation analysis revealed that these mutations were allelic to *kan-2* (ref. 5). Alleles *kan-11* and *kan-15* were generated in Col using ethylmethane sulphonate, and *kan-12* and *kan-14* were generated in Ler and Col genetic backgrounds, respectively, by ionizing radiation. The *kan-13* line, generated in Ws by T-DNA insertion mutagenesis, was produced by K. Feldman and obtained from the *Arabidopsis* Biological Resource Center. All six alleles are phenotypically very similar, with the exception of *kan-14*, which displays an additional serrated leaf phenotype. *kan-12* was crossed to Col eight times before phenotypic characterization.

Subcellular localization

KAN–GUS and KAN–mGFP5 fusions were constructed by amplifying the KAN open reading frame (ORF) from cDNA clone A1-2 using the following oligonucleotides: tggccATGgCTATGGAAGGTGTTTTC, which converts the first ATG into an NcoI site, and tctagatcTTTCTCGTGCCCAATCTGGTC, which removes the stop codon, introduces a BglII site and aligns the reading frame of KAN with GUS in pCambia1381Z and mGFP5 in pCambia1302 (ref. 21)—lowercase letters indicate nucleotide mismatches with the KAN sequence. The PCR product was digested and cloned into the NcoI and BglII sites in the vectors. Qiagen-purified plasmid DNA was precipitated on 1.6 µm gold microcarriers (BioRad catalogue number 1652264), transferred to macrocarriers (1652335), introduced into onion adaxial epidermal peels using a BioRad Biolistic PDS-1000/He System with 1,100 pounds-per-square-inch rupture disks (1652329) at 710 mm Hg of vacuum and a target distance of 6 cm.

Histology and expression analysis

Tissue fixation, paraffin embedding, sectioning and *in situ* hybridization of digoxigenin-UTP labelled RNA probe to mRNA was performed as described²². Anti-sense RNA probe was transcribed with T7 RNA polymerase from the full-length A1-2 cDNA clone after linearizing with BamHI. Sense control probes were transcribed with T3 RNA polymerase using the same clone linearized with Asp718. Tissues for histological examination were fixed, embedded in Spurr's resin, sectioned and stained as described²³. Measurements of cell area were performed on a Macintosh computer using NIH Image (<http://rsb.info.nih.gov/nih-image/>). We prepared camera lucida drawings using whole mounts of mature leaves vacuum-infiltrated with water.

Transgenic plants

The KAN misexpression construct was derived from the cDNA clone A1-2. 352 bp of the 5' untranslated leader was removed by digesting with BamHI, blunting with mung bean exonuclease, partially digesting with XmnI and re-ligating to generate pDERT4. pDERT4 was digested with NotI to free the coding region, overhangs were filled in with Klenow and dNTPs, and the fragment was cloned into the SmaI site of pROK2. Construct orientation was confirmed by sequencing the junctions between the KAN coding region, the CAMV 35S promoter, and the nopaline synthase terminator. The construct was introduced into *Agrobacterium tumefaciens* strain GV3101 pMP90 (ref. 24) by electroporation (2.5 v, 400 Ω). *Arabidopsis* plants were transformed using the floral dip method¹⁸, and seedlings were selected on plates (1/2× Gibco MS salts, 3% bactoagar, 75 mg l⁻¹ kanamycin sulphate).

Received 22 December 2000; accepted 18 April 2001.

- Telfer, A. & Poethig, R. S. In *Arabidopsis* (eds Meyerowitz, E. M. & Somerville, C. R.) 379–401 (Cold Spring Harbor Laboratory Press, Plainview, 1994).
- Chien, J. C. & Sussex, I. M. Differential regulation of trichome formation on the adaxial and abaxial leaf surfaces by gibberellins and photoperiod in *Arabidopsis thaliana* (L.) Heynh. *Plant Physiol.* **111**, 1321–1328 (1996).
- Telfer, A., Bollman, K. M. & Poethig, R. S. Phase change and the regulation of trichome distribution in *Arabidopsis thaliana*. *Development* **124**, 645–654 (1997).
- Bowman, J. L. & Smyth, D. R. CRABS CLAW, a gene that regulates carpel and nectary development in *Arabidopsis*, encodes a novel protein with zinc finger and helix-loop-helix domains. *Development* **126**, 2387–2396 (1999).
- Eshed, Y., Baum, S. F. & Bowman, J. L. Distinct mechanisms promote polarity establishment in carpels of *Arabidopsis*. *Cell* **99**, 199–209 (1999).
- Behringer, F. J. & Medford, J. I. A plasmid rescue technique for the recovery of plant DNA disrupted by T-DNA insertion. *Plant Mol. Biol. Rep.* **10**, 190–198 (1992).
- Kieber, J. J., Rothenberg, M., Roman, G., Feldmann, K. A. & Ecker, J. R. CTR1, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the raf family of protein kinases. *Cell* **72**, 427–441 (1993).
- Riechmann, J. L. *et al.* *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105–2110 (2000).
- Hall, L. N., Rossini, L., Cribb, L. & Langdale, J. A. GOLDEN 2: a novel transcriptional regulator of cellular differentiation in the maize leaf. *Plant Cell* **10**, 925–936 (1998).
- Sakai, H., Aoyama, T., Bono, H. & Oka, A. Two-component response regulators from *Arabidopsis thaliana* contain a putative DNA-binding motif. *Plant Cell Physiol.* **39**, 1232–1239 (1998).
- Wykoff, D. D., Grossman, A. R., Weeks, D. P., Usuda, H. & Shimogawara, K. Psr1, a nuclear localized protein that regulates phosphorus metabolism in *Chlamydomonas*. *Proc. Natl Acad. Sci. USA* **96**, 15336–15341 (1999).
- Sakai, H., Aoyama, T. & Oka, A. *Arabidopsis* ARR1 and ARR2 response regulators operate as transcriptional activators. *Plant J.* **24**, 703–711 (2000).
- Imamura, A. *et al.* Compilation and characterization of *Arabidopsis thaliana* response regulators implicated in His-Asp phosphorelay signal transduction. *Plant Cell Physiol.* **40**, 733–742 (1999).
- Lohrmann, J. *et al.* Differential expression and nuclear localization of response regulator-like proteins from *Arabidopsis thaliana*. *Plant Biol.* **1**, 495–505 (1999).
- Makino, S. *et al.* Genes encoding pseudo-response regulators: insight into His-to-Asp phosphorelay and circadian rhythm in *Arabidopsis thaliana*. *Plant Cell Physiol.* **41**, 791–803 (2000).
- Blom, N., Gammeltoft, S. & Brunak, S. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362 (1999).
- Varagana, M. J., Schmidt, R. J. & Raikhel, N. V. Nuclear localization signal(s) required for nuclear targeting of the maize regulatory protein Opaque-2. *Plant Cell* **4**, 1213–1227 (1992).
- Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
- Lynn, K. *et al.* The PINHEAD/ZWILLE gene acts pleiotropically in *Arabidopsis* development and has overlapping functions with the ARGONAUTE1 gene. *Development* **126**, 469–481 (1999).
- McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).

- Roberts, C. S. *et al.* in *pCambia Vector Release Manual Version 3.05* (Cambia, Canberra, 1997).
- Lincoln, C., Long, J., Yamaguchi, J., Serikawa, K. & Hake, S. A knotted1-like homeobox gene in *Arabidopsis* is expressed in the vegetative meristem and dramatically alters leaf morphology when overexpressed in transgenic plants. *Plant Cell* **6**, 1859–1876 (1994).
- Barton, M. K. & Poethig, R. S. Formation of the shoot apical meristem in *Arabidopsis thaliana*: an analysis of development in the wild type and shoot meristemless mutant. *Development* **119**, 823–831 (1993).
- Koncz, C. & Schell, J. The promoter of T₁-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Mol. Gen. Genet.* **204**, 383–396 (1986).
- Patharkar, O. R. & Cushman, J. C. A stress-induced calcium-dependent protein kinase from *Mesembryanthemum crystallinum* phosphorylates a two-component pseudo-response regulator. *Plant J.* **24**, 679–691 (2000).

Acknowledgements

We thank Y. Eshed and J. Bowman for pointing out the similarity between *kanadi* and our *early trichomes* mutant and for providing seeds of *kan-2* for allelism tests. We also thank K. Barton for suggesting that KAN might be involved in radial as well as adaxial/abaxial patterning and for providing some of the tissue used for *in situ* hybridization; M. Aukerman for the pistil scanning electron micrograph; and M. Bucan for helpful comments on this manuscript. This work was supported by NIH Postdoctoral Research Fellowship (R.K.) and by a DOE grant (R.S.P.).

Correspondence and requests for materials should be addressed to R.S.P. (e-mail: spoethig@sas.upenn.edu). The cDNA sequence for KAN is deposited in GenBank under accession number AY030192.

Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots

Jane R. McConnell^{†‡}, John Emery[§], Yuval Eshed[§], Ning Bao^{*||}, John Bowman[§] & M. Kathryn Barton^{*}

^{*} Department of Genetics, [†] Program in Molecular and Cellular Biology and ^{||} Program in Plant Breeding and Plant Genetics, University of Wisconsin–Madison, 445 Henry Mall, Madison, Wisconsin 53706, USA
[§] Section of Plant Biology, University of California–Davis, One Shields Avenue, Davis, California 95616, USA

The upper side of the angiosperm leaf is specialized for efficient capture of sunlight whereas the lower side is specialized for gas exchange. In *Arabidopsis*, the establishment of polarity in the leaf probably requires the generation and perception of positional information along the radial (adaxial versus abaxial or central versus peripheral) dimension of the plant. This is because the future upper (adaxial) side of the leaf develops from cells closer to the centre of the shoot, whereas the future under (abaxial) side develops from cells located more peripherally. Here we implicate the *Arabidopsis* PHABULOSA and PHAVOLUTA genes in the perception of radial positional information in the leaf primordium. Dominant *phabulosa* (*phb*)¹ and *phavoluta* (*phv*) mutations cause a dramatic transformation of abaxial leaf fates into adaxial leaf fates. They do so by altering the predicted sterol/lipid-binding domains of ATHB14 and ATHB9, proteins of previously unknown function that also contain DNA-binding motifs. This change probably renders the protein constitutively active, implicating this domain as a central regulator of protein function and the PHB and PHV proteins as receptors for an adaxializing signal.

Dominant *phabulosa*¹ and *phavoluta* mutations (Fig. 1) cause transformation of abaxial to adaxial fates in leaves and leaf-like organs; the most severely affected organs develop with radial

[‡] Present address: Carnegie Institute of Washington, Department of Plant Biology, Stanford, California 94305, USA.

symmetry and exhibit adaxial traits around their circumference. *phabulosa* and *phavoluta* mutations map to the *ATHB14* and *ATHB9* genes, respectively (see Methods). *ATHB9* and *ATHB14* are members of a plant-specific class of homeodomain–leucine zipper (HD–ZIP)–containing proteins² (Fig. 2). The HD–ZIPIII protein subtype, which includes *ATHB8*, *ATHB9*, *REVOLUTA*/INTERFASCICULAR FIBERLESS^{3,4,5} and *ATHB14*, is characterized by an amino-terminal HD–ZIP motif followed by a region with similarity to a mammalian sterol/lipid-binding domain (START domain)⁶. If the predicted START domain in these family members has a regulatory function, we expect to find mutations that disrupt this regulation. Such mutations are found in all five dominant *phb* and all four dominant *phv* alleles.

Both *phb-1d* and *phb-2d* contain a guanine to adenine change at the first position of intron 4 (Fig. 2). This intron is in the region coding for the START domain. Analysis of polymerase chain reaction after reverse transcription (RT-PCR) products made from *phb-1d* homozygotes show that this mutation affects splicing of the *ATHB14* transcript. In addition to normally spliced transcripts (10 out of 24 RT-PCR-derived clones), transcripts that retained intron 4 (2 out of 24) and transcripts that arose from the use of a cryptic splice donor +30 (11 out of 24) or +33 (1 out of 24) within intron 4 were found. The latter transcripts cause 10 and 11 amino-acid in-frame insertions, respectively, in the N-terminal portion of the START domain.

To determine which, if any, of the altered transcripts causes the Phb phenotype, a wild-type *ATHB14* complementary DNA, a mutant *ATHB14* cDNA carrying the 33-residue insertion and a truncated cDNA ending at the first stop in intron 4 (the product expected from transcripts that retain intron 4) were each fused to the constitutive cauliflower mosaic virus (CAMV) 35S promoter. These constructs were transformed into wild-type *Arabidopsis* plants. Transformants carrying the wild-type cDNA ($n = 12$) or the truncated cDNA ($n = 9$) developed leaves with normal polarity.

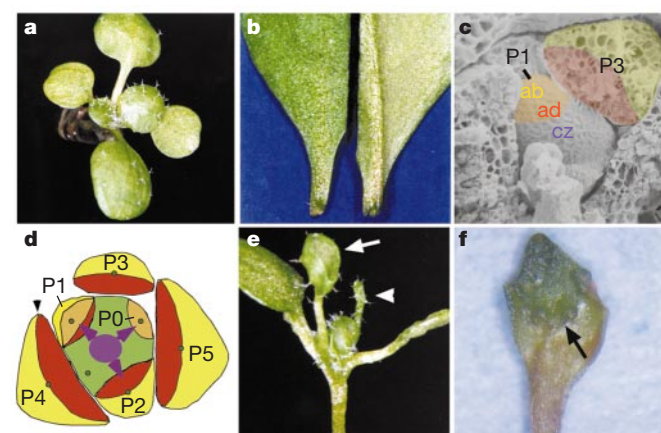


Figure 1 Development of adaxial and abaxial leaf domains. **a**, Wild-type *Arabidopsis* plant. Adaxial leaf surfaces face up. **b**, Adaxial (left) and abaxial (right) leaf surfaces. **c**, Scanning electron micrograph of a vegetative SAM. For stages of leaf development see ref. 14. The adaxial domain of P3 is red; the abaxial domain is yellow. Approximate boundaries of P1 are shown in orange. The central zone (cz), adaxial (ad) and abaxial leaf domains (ab) are at increasing distances from the centre of the plant. **d**, Model for the acquisition of leaf polarity. P0 (orange) lacks overt molecular polarity^{7,9}. Adaxial (red) and abaxial (yellow) develop in response to an unequally graded signal, perhaps emanating from the centre of the plant^{10,11}. Once established, juxtaposition of adaxial and abaxial leaf fates (arrowhead) promotes outgrowth of the leaf blade¹⁵. Green circles indicate vasculature. **e**, *phv-1d* mutant. Note rod-shaped (arrowhead) and trumpet-shaped (arrow) leaves with adaxial characters around their circumference. **f**, Underside of leaf expressing an *ATHB14* cDNA containing the 33-bp insertion in the START domain. The distal part of the leaf exhibits adaxial traits (arrow).

Failure of these plants to show a phenotype indicate that *ATHB14* alone is insufficient to direct ectopic adaxial development, consistent with *ATHB14* requiring a ligand for function. In contrast, half the transformants carrying the cDNA with the 33-base-pair (bp) insertion (25 out of 53) developed leaves with adaxialized phenotypes similar to those seen in *phb-1d* mutants (Fig. 1f). We conclude that the insertion of a short stretch of amino acids in the START domain of *ATHB14* causes the Phb phenotype.

The remaining *phb* alleles (*phb-3d*, *-4d* and *-5d*) cause a glycine to glutamic acid change one amino acid away from the site of the insertion in *phb-1d* (Fig. 2). Of all the *ATHB* genes, *ATHB9* has the highest sequence similarity to *ATHB14* (85% amino-acid identity throughout the protein; Fig. 2). All four *phv* alleles (*phv-1*, *-2*, *-3* and *-4*) cause a substitution of glycine to glutamic acid in *ATHB9* at the site corresponding to the same substitution in *phb-3d* in *ATHB14* (Fig. 2). The glycine altered in these mutants is at a site occupied by a polar residue in the START consensus sequence⁶.

The repeated isolation of just two types of alterations and their proximity indicates that only a small target in the *PHB* and *PHV* genes mutate to yield a Phb phenotype. Moreover, the adaxialized phenotype caused by such mutations in both *PHABULOSA* and *PHAVOLUTA* indicates their proteins have similar functions.

The *PHB* transcript is preferentially expressed in the adaxial domain of the developing leaf (Fig. 3). It is found throughout the P0 leaf primordium (Fig. 3b–d). Transcript levels increase in P1 and become preferentially localized to the adaxial leaf domain by the P2 stage (Fig. 3a–d). Expression throughout the entire P0 primordium followed by the polar accumulation of transcript by the P2 stage of leaf development has been observed for the adaxially localized *PINHEAD*⁷ and the abaxially localized *YABBY* transcripts^{8,9}. Thus, when first specified, the leaf primordium lacks molecular evidence of polarity and only later becomes a polarized entity with adaxial and abaxial transcripts separated into the appropriate domains.

Embryonic *PHB* expression parallels expression in the vegetative

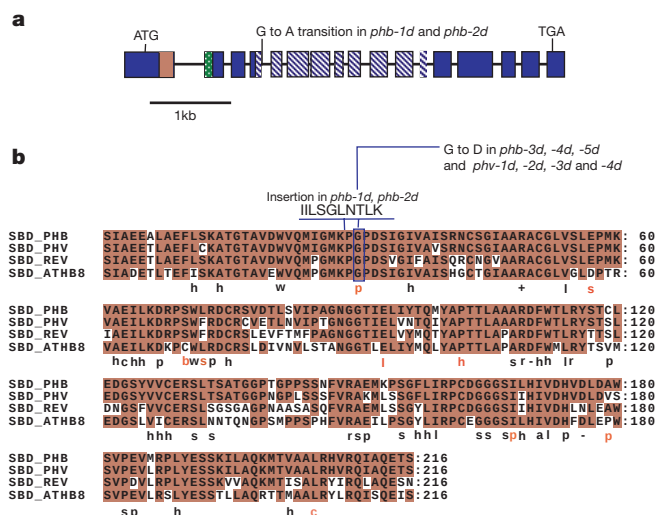


Figure 2 Dominant *phb* and *phv* mutations alter the START domain. **a**, Organization of the *ATHB14/PHB* gene. Red, homeodomain; green, leucine zipper; hatched, predicted START domain. Site of mutation in *phb-1d* and *phb-2d* is shown. **b**, Alignment of the START domain of the *PHABULOSA* (*ATHB14*), *PHAVOLUTA* (*ATHB9*), *REVOLUTA* and *ATHB8* proteins. Shaded residues are identical to *PHB*. Insertion in *phb-1d* and *phb-2d* is caused by use of a cryptic donor site at +30 or +33 in intron 4. Consensus residues for the START domain⁶ are shown below alignment. Sites at which *PHB* does not agree with consensus are shown in red letters. H, hydrophobic; w, tryptophan; p, polar; +, positive charge; l, aliphatic; s, small; c, charged; b, big; -, negative charge; r, arginine.

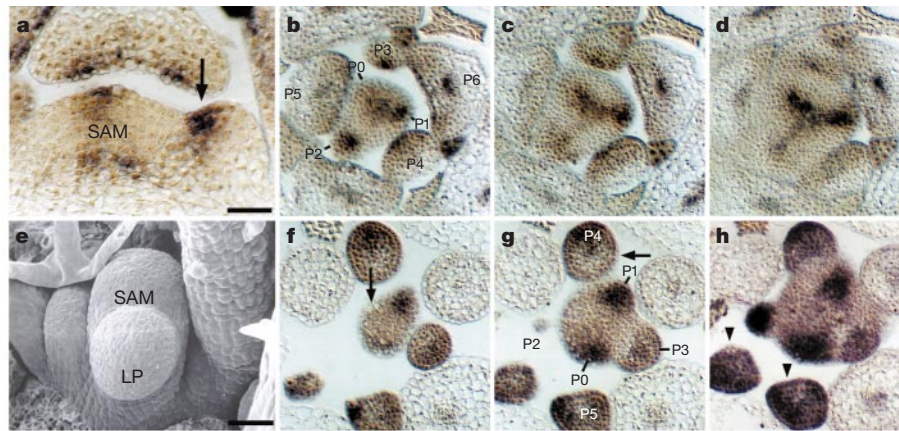


Figure 3 *PHB* expression in wild-type and *phb-1d* SAMs. **a**, Longitudinal section through a wild-type SAM showing *PHB* expression in the centre of the meristem and on the adaxial side of the P2 leaf (arrow). **b–d**, Transverse serial sections through a wild-type SAM. Section **b** is uppermost. *PHB* RNA becomes preferentially localized to the adaxial leaf domain. Expression is also observed in the developing vasculature. **e**, Scanning electron micrograph of a *phb-1d* meristem. LP, leaf primordium. **f–h**, Transverse serial sections through a *phb-1d* SAM. **f**, Transverse section about 10 μ m into the meristem. The

meristem is indicated by an arrow. Primordia are radial in cross section. **g**, *PHB* can accumulate more in the abaxial domain than in the adaxial domain of the primordia (arrow). **h**, Transverse section at about 21 μ m below the SAM summit. Arrowheads point to two primordia that express *PHB* throughout the adaxial and abaxial domains. Scale bars: **a–d**, **f**, **g**, 50 μ m; **e**, 30 μ m. Hybridizations were performed as described in ref. 15. The probe corresponds to nucleotides 1,376–2,297 of the *PHB* coding sequence.

shoot apical meristem (SAM) (Fig. 4). *PHB* transcript is initially present throughout the presumptive cotyledons and only later (globular stage just before cotyledon outgrowth) is seen at high levels in their adaxial domains.

The adaxial location of *PHB* indicates that wild-type *PHB* specifies adaxial leaf fates and that the dominant lesions in the *PHB* and *PHV* mutants cause these gene products to be constitutively active. Supporting this hypothesis, *PHB* is expressed on both sides of leaf and cotyledon primordia in *phb-1d* mutants (Fig. 3e–h), in some cases appearing more intense in the abaxial domain. Levels of *PHB* transcript are also higher in mutant than in wild-type tissue; less than 2 h of incubation with the chromogenic substrate was required to see signal in the mutant, whereas overnight incubation was required for signal to develop to equivalent levels in the wild type. Because the primary lesion in the *phb-1d* mutant is at the level of the gene product, the ectopic and overexpression of *PHB* may be due to positive feedback of active *PHB* on *PHB* transcript levels and location.

Gradients of *PHB* accumulation can be seen in wild-type embryos. The highest levels of *PHB* RNA are in adaxial (or central) regions of the embryo (Fig. 4). Three to four distinct colour values (indicating varying expression levels) are seen in these gradients. A similar gradient may also be present in the meristem (Fig. 3a), with highest expression at the meristem summit. Such gradients are consistent with a role for *PHB* in defining and/or interpreting positional identity along the plant's radial axis. Moreover, if the positive feedback of activated *PHB* on *PHB* transcript levels occurs in wild-type plants, then high levels of *PHB* transcript should reflect regions of high *PHB* activity, whereas low levels should reflect regions of low *PHB* activity. This, in turn, should reflect an underlying graded distribution of an activator (or repressor) of *PHB* activity, perhaps the ligand for the START domain.

The adaxial domains of developing leaves together with the meristem comprise a roughly cylindrical central domain based on their shared pattern of *PNH* expression⁷. This implies that positional information used to distinguish central from peripheral features is also used to distinguish the adaxial and abaxial halves of the leaf. A new feature of *PHB* expression strengthens the idea that the adaxial leaf domain and the meristem behave as a unit. Although *PHB* expression is low throughout the SAM, higher-level expression extends in rays from the centre of the SAM to the P0- and P1-stage primordia, giving the meristem the appearance of a clock

face (Fig. 3a–d). Thus, in young primordia, the unit accumulating *PHB* transcript is the adaxial leaf domain and adjacent meristem cells. Only a weak remnant of *PHB* expression in the meristem adjacent to P2 primordia is visible (Fig. 3b–d) and by P3, *PHB* expression in the adjacent meristem is gone. Similar patterns are seen in the embryo: *PHB* RNA is found in the young cotyledons and adjacent meristem, disappears in the meristem as the cotyledons develop, but returns to the SAM at the walking-stick stage (Fig. 4) when the first two true leaves begin to develop in the SAM (J. A. Long and M.K.B., unpublished observations). Notably, the rays of *PHB* expression in the meristem are missing in *phb-1d* mutants (Fig. 3f–h); their formation may depend on the development of polar leaf primordia and/or the presence of abaxial leaf domains.

The *phb-1d* mutation is one of a few mutations that cause formation of ectopic meristems—*phb-1d* mutants form ectopic meristems from the undersides of their leaves¹. Axillary meristems normally develop from the adaxial leaf base. One explanation for the ectopic meristems is that *phb-1d* causes the development of extra-adaxial leaf-base tissue, which in turn induces a meristem. Alternatively, the *phb-1d* mutation promotes the development of both meristem and adaxial leaf in parallel. The latter hypothesis is supported by the 'unit' of *PHB* expression—adaxial leaf and adjacent meristem—which is seen in early leaf development.

We propose the following model for *PHB* in the development of leaf polarity. Low levels of *PHB* protein are present throughout the young, unpolarized P0 leaf primordium. As the leaf develops, a ligand for *PHB* is unequally distributed throughout the primordium, with highest levels present in the cells closest to the centre of the meristem. This ligand activates the *PHB* protein. (Alternatively, one can postulate a negatively acting ligand with highest concentration in the abaxial domain.) Active *PHB* promotes adaxial leaf development and also positively feeds back on the synthesis and/or stability of its own transcript. This feedback loop results in a persistent, adaxially located domain of *PHB* synthesis. To maintain *PHB* activity, this feedback loop could also positively regulate ligand synthesis and/or stability. Such a feedback loop would create a new reference point for adaxial position for the developing leaf as it grows away from its original source of positional information. This is consistent with the finding that, after a certain stage in early primordium development, polarity was no longer abolished by surgical isolation of the primordium^{10,11}. In dominant *phb* and *phv* mutants, *PHB* and *PHV* proteins, respectively, are constitutively

active by virtue of changes in the predicted START domain; ligand is no longer required to activate the proteins. (If the ligand were negatively acting, it would no longer be able to repress PHB action.) This results in high levels of PHB activity in both domains of the primordium and causes them to develop adaxial characteristics. Loss of *PHB* function is predicted to result in an abaxialized phenotype. Given the high degree of similarity and the evidence for functional interchangeability of *PHABULOSA* and *PHAVOLUTA*, it

is likely that inactivating both genes will be required to see a loss-of-function phenotype. Of note, recessive mutations in the related *REVOLUTA* gene result in narrow leaves and a lack of axillary meristems⁵. The *Rev* phenotype is consistent with the disruption of the adaxial leaf domain and its requirement for axillary meristem formation. □

Methods

Genetics

The *phb-1d* mutation is incompletely dominant and is also temperature sensitive¹. Unless otherwise noted, experiments were carried out on *phb-1d* homozygotes grown at restrictive temperature (24 °C). Additional mutants with dominant *Phb* phenotypes were isolated in three independent screens of M2 plants. M1 seed was subjected to mutagenesis with 0.175% ethylmethane sulphonate for 12 h. M2 seeds from roughly 4,000 M1 plants were screened. Four mutants carry mutations in the *PHABULOSA* gene (*phb-2d*, *-3d*, *-4d* and *-5d*) and four carry mutations in the *ATHB9* gene. Because of the molecular similarity of *ATHB9* to both the *PHABULOSA* and *REVOLUTA* genes and the phenotypic similarity of mutants in *ATHB9* to *phb* mutants, we have named the *ATHB9* gene *PHAVOLUTA* and designate the mutations *phv-1d*, *-2d*, *-3d* and *-4d*.

The *phb-1d* mutation lies between *argonaute-like* (at 13,921 kb; 17 out of 852 recombinant chromosomes) and a *phantastica* homologue (at 17,531 kb; 29 out of 830 recombinant chromosomes) on chromosome 2. No recombination was observed between *phb-1d* and markers at positions 14,561 or 14,653 kb (0 out of 430 recombinant chromosomes). The *ATHB14* gene is at position 14,588 kb/67.9 cM. *phb-2d* is linked to marker nga361 (position 63 cM; 3 out of 36 recombinant chromosomes). For *phb-4d*, the mutant phenotype was mapped relative to the lesion in *ATHB14* that causes the glycine to glutamic acid change (the *phb-4d* mutation itself). The lesion was detected using primers (5'-CCAAATCCTCAGCATCAGCA-3' and 5'-CGCAAGGACGGATGAGAAAC-3') to amplify a 894-bp fragment surrounding the beginning of exon 5 of *ATHB14*. The *phb-4d* mutation destroys a *ScrFI* site at the beginning of this exon, thus generating a scorable polymorphism. No recombinants (0 out of 30) were detected. The same strategy was used for mapping *phv-1d*, which destroys a *ScrFI* site near the beginning of exon 5 in *ATHB9* except that the primers used were 5'-GCAAAACCCAACATCAGC-3' and 5'-CACAAGGACGGATAAGAAAC-3'.

Analysis of *phb-1d* transcripts

RT-PCR products spanning the splice junctions of the fourth intron of *ATHB14* were made from RNA isolated from *phb-1d* homozygotes grown at 22 °C. RNA was isolated from homozygous *phb-1d* individuals using the RNeasy Kit (Qiagen). The reverse transcription reaction was performed using avian myeloblastosis virus reverse transcriptase (Promega). The reverse transcription products served as template for three independent PCR reactions using EX-Taq as the polymerase (Panvera). Products from the PCR reactions were cloned into the pGEMT Easy Vector (Promega), confirmed by digestion, and subsequently sequenced. We sequenced eight clones from each PCR reaction. As a control, wild-type messenger RNA was isolated, reverse transcribed and sequenced in a side-by-side analysis. The experiment was repeated three times with similar results.

Transformation experiments

A full-length and a truncated (the first 600 bp) *PHB* cDNA were isolated from the *Arabidopsis thaliana* Stock Centre CD4-16 cDNA (WS ecotype) library using the following gene specific primers: 5'-TTGGTACCATGATGATGGTCCATTCGATGAGC-3' (5' primer), 5'-TGAGCTCTCAAACGACGACCAATTCACGAA-3' (3' primer for full-length cDNA) and 5'-TGAGCTCTCTCATCCCAATCATCTGACCCAGTC-3' (3' primer for the truncated cDNA). *KpnI* and *SacI* sites were added to ends of the primers and the amplified products were cloned into pAF21 (Anita Fernandez), which contains both the CAMV 35S promoter¹² and a nopaline synthase terminator. A full-length *ATHB14* cDNA containing the 33-bp insertion was generated using RT-PCR on RNA isolated from *phb-1d* homozygotes. The DNA sequence of all clones was determined to rule out amplification and/or cloning errors. Plant transformation was by the floral dip method¹³. Seeds were collected, sterilized and plated to MS medium containing 40 µg ml⁻¹ hygromycin to select for transformants.

Received 16 January; accepted 17 April 2001.

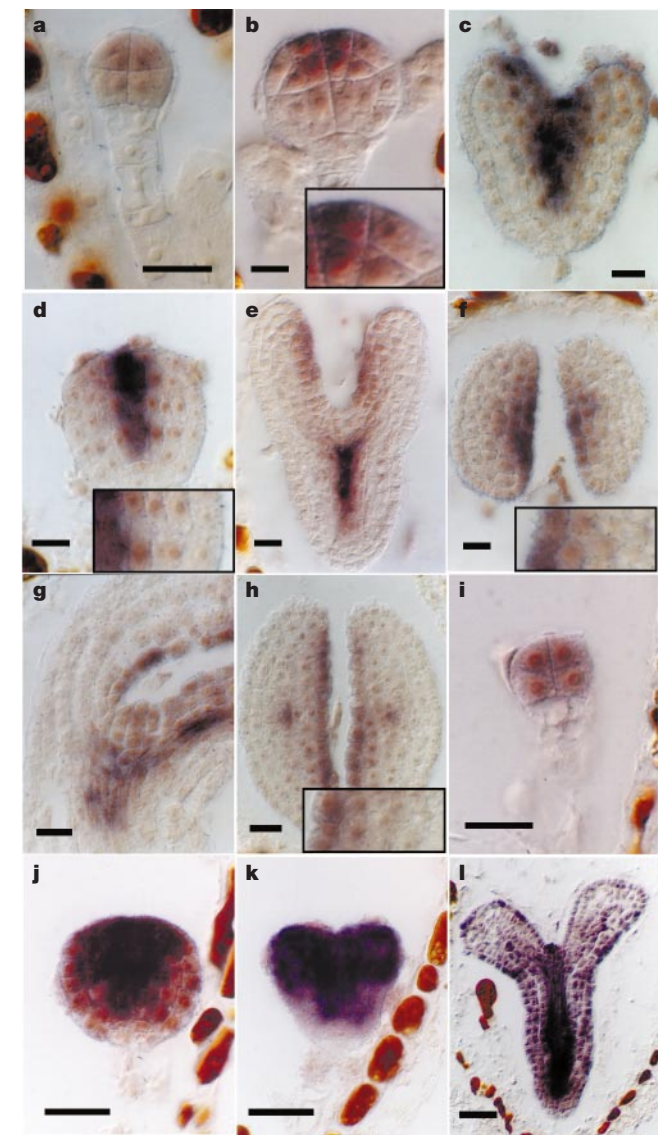


Figure 4 Expression of *PHB* in embryos. **a–h**, Wild-type embryos. **i–l**, Homozygous *phb-1d* embryos. **a**, A 16-cell-stage embryo expresses *PHB* in all cells of the embryo proper but not in the suspensor or hypophysis. **b**, In a 32-cell-stage embryo, expression is stronger in the top half. Insets, gradients of transcript accumulation with highest levels in adaxial or central regions. **c**, Heart-stage embryo. Expression is strongest in the most central regions of the embryo, including the presumptive meristem and the adaxial regions of the cotyledons. **d**, Heart-stage embryo in sagittal section. **e**, Torpedo-stage embryo retains *PHB* expression in the vascular cylinder and in the adaxial regions of the cotyledons but has much reduced expression in the developing SAM. **f**, Section through cotyledons of a torpedo-stage embryo. **g**, Embryo at walking-stick stage (*PHB* expression has returned in the SAM). **h**, Section through cotyledons of walking-stick embryo. **i**, *phb-1d* embryo. *PHB* RNA is found throughout the early-globular-stage embryo. **j**, Late-stage *phb-1d* globular embryo; *PHB* RNA levels are higher in both the top and bottom halves of the embryo than in wild-type embryos. **k**, Heart-stage *phb-1d* embryo. Both adaxial and abaxial domains of the presumptive cotyledons accumulate *PHB*. **l**, *phb-1d* torpedo-stage embryo. *PHB* RNA is found throughout the embryo. Scale bars: **a–k**, 25 µm; **l**, 50 µm.

- McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).
- Sessa, G., Steindler, C., Morelli, G. & Roberti, I. The *Arabidopsis* *Atb-8*, *-9* and *-14* genes are members of a small gene family coding for highly related HD-ZIP proteins. *Plant Mol. Biol.* **38**, 609–622 (1998).
- Zhong, R. & Ye, Z.-Y. *IFL1*, a gene regulating interfascicular fiber differentiation in *Arabidopsis*, encodes a homeodomain-leucine zipper protein. *Plant Cell* **11**, 2139–2152 (1999).
- Ratcliffe, O. J., Riechmann, J. L. & Zhang, J. Z. *INTERFASCICULAR FIBERLESS1* is the same gene as *REVOLUTA*. *Plant Cell* **12**, 315–317 (2000).
- Talbert, P., Adler, H. T., Parks, D. W. & Comai, L. The *REVOLUTA* gene is necessary for apical meristem development and for limiting cell divisions in the leaves and stems. *Development* **121**, 2723–2735 (1995).
- Ponting, C. P. & Aravind, L. START: a lipid-binding domain in *StAR*, HD-ZIP and signalling proteins. *Trends Biochem. Sci.* **24**, 130–132 (1999).
- Lynn, K. et al. The *PINHEAD/ZWILLE* gene acts pleiotropically in *Arabidopsis* development and has overlapping functions with the *ARGONAUTE1* gene. *Development* **126**, 469–481 (1999).
- Sawa, S. et al. *FILAMENTOUS FLOWER*, a meristem and organ identity gene of *Arabidopsis*, encodes a protein with a zinc finger and HMG-related domains. *Genes Dev.* **13**, 1079–1088 (1999).
- Siegfried, K. R. et al. Members of the *YABBY* gene family specify abaxial cell fate in *Arabidopsis*. *Development* **126**, 4117–4128 (1999).

10. Sussex, I. M. Morphogenesis in *Solanum tuberosum* L: experimental investigation of leaf dorsiventrality and orientation in the juvenile shoot. *Phytomorphology* **5**, 286–300 (1955).
11. Snow, M. & Snow, R. The dorsiventrality of leaf primordia. *New Phytol.* **8**, 188–207 (1959).
12. Benfey, P. N., Ren, L. & Chua, N. H. Tissue-specific expression from CaMV 35S enhancer subdomains in early stages of plant development. *EMBO J.* **9**, 1677–1684 (1990).
13. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
14. Long, J. A. & Barton, M. K. Initiation of axillary and floral meristems in *Arabidopsis*. *Dev. Biol.* **125**, 341–353 (2000).
15. Waites, R. & Hudson, A. *phantastica*: a gene required for dorsoventrality of leaves in *Antirrhinum majus*. *Development* **121**, 2143–2154 (1995).

Acknowledgements

We would like to thank M. Evans and K. Lynn for comments on the manuscript and A. Hamilton, A. Derr and B. Bordini for technical assistance. This work was supported by grants from the Beckman Foundation, NSF and the USDA.

Correspondence and requests for materials should be addressed to M.K.B. (e-mail: mkbarton@facstaff.wisc.edu).

Defects in mismatch repair promote telomerase-independent proliferation

Aylin Rizki* & Victoria Lundblad*†

* Departments of Biochemistry and Molecular Biology, and † Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

Mismatch repair has a central role in maintaining genomic stability by repairing DNA replication errors and inhibiting recombination between non-identical (homeologous) sequences^{1,2}. Defects in mismatch repair have been linked to certain human cancers, including hereditary non-polyposis colorectal cancer (HNPCC) and sporadic tumours^{3–5}. A crucial requirement for tumour cell proliferation is the maintenance of telomere length⁶, and most tumours achieve this by reactivating telomerase⁷. In both yeast and human cells, however, telomerase-independent telomere maintenance can occur as a result of recombination-dependent exchanges between often imperfectly matched telomeric sequences^{8–12}. Here we show that loss of mismatch-repair function promotes cellular proliferation in the absence of telomerase. Defects in mismatch repair, including mutations that correspond to the same amino-acid changes recovered from HNPCC tumours¹³, enhance telomerase-independent survival in both *Saccharomyces cerevisiae* and a related budding yeast with a degree of telomere sequence homology that is similar to human telomeres. These results indicate that enhanced telomeric recombination in human cells with mismatch-repair defects may contribute to cell immortalization and hence tumorigenesis.

The mismatch-repair (MMR) pathway contains at least five genes in yeast—*MSH2*, *MSH3*, *MSH6*, *MLH1* and *PMS1*—which are needed to avoid mutation and repress homeologous recombination^{1,2}. Potential natural targets of this recombination fidelity mechanism might be the telomeric and subtelomeric regions of chromosomes, which are not perfectly homologous in either yeast or human cells^{14,15}. As recombination between chromosome ends is a means by which cells can maintain long-term proliferation in the absence of telomerase^{8–12}, the anti-recombination activity of the MMR machinery might potentially have an inhibitory effect on the ability of telomerase-defective cells to proliferate.

To test this possibility, we examined the effects of MMR mutations on telomerase-independent survival in budding yeast. Yeast strains lacking telomerase exhibit healthy growth initially, followed by a rapid decline in colony formation by ~75 generations¹⁶ (Fig. 1a). Because telomerase-defective strains that are late in their senescence progression display notoriously variable growth characteristics (refs 8, 9, 16; and see below), we used several complementary assays to assess the progressive decline in viability of an *est2-Δ* strain (deleted for the catalytic subunit of telomerase¹⁷) and an *est2-Δ msh2-Δ* double-mutant strain.

First, we evaluated the growth characteristics of multiple samples of each genotype after propagation by successive colony formation for up to ~100 generations (selected examples are shown in Fig. 1a). A genotype-blind comparison of the relative growth characteristics of 16 *est2-Δ* samples and 18 *est2-Δ msh2-Δ* samples, collected at four time points during outgrowth, showed that the *est2-Δ msh2-Δ* strain exhibited a less pronounced senescence phenotype at both ~75 and ~100 generations (Fig. 1b).

This relative growth comparison was supported by a more quantitative measurement, in which dilutions of a fixed quantity of cells recovered after propagation for ~25 to 100 generations were scored for their ability to grow (Fig. 1c). Data collected in this manner, from nine *est2-Δ msh2-Δ* samples and nine *est2-Δ* samples, gave results that were indistinguishable from those shown in Fig. 1b (data not shown). Comparison of growth curves from three independent cultures of each genotype also supported the conclusion that an *est2-Δ msh2-Δ* strain showed improved growth when compared with the more rapidly senescing *est2-Δ* strain (Fig. 1d).

An additional sensitive assay showed that loss of *MSH2* function enhances the proliferation of a strain lacking telomerase even at a very early time point. Cells from *est2-Δ msh2-Δ* and *est2-Δ* strains early in their senescence progression were mixed at a 50:50 ratio, grown in continuous log phase, and sampled at ~10-generation intervals for the relative contribution of each strain to the culture (Fig. 1e). In seven independent experiments, the *est2-Δ msh2-Δ* strain displayed a competitive growth advantage relative to the *est2-Δ* strain after ≤ 50 generations of mixed culture growth. This was not due to an inherent growth advantage of *msh2-Δ* strains (Fig. 1e, right). Therefore, even at a time point when *est2-Δ* and *est2-Δ msh2-Δ* strains could not be distinguished readily by their colony growth properties (~35 generations), the *est2-Δ msh2-Δ* strain had a clear growth advantage. Thus, the effects due to loss of *MSH2* are rapid, arguing that growth enhancement is not a secondary consequence of suppressor mutations (see also below).

This growth enhancement was not due to an effect of the absence of *MSH2* on telomere length, as the average rate of telomere attrition of four individual telomeres was comparable in *est2-Δ* and *est2-Δ msh2-Δ* strains (3.40 ± 0.9 versus 3.38 ± 1.2 base pairs per generation, respectively). In addition, isolates recovered from *est2-Δ msh2-Δ* cultures that had re-acquired healthy growth exhibited both classes of the characteristic telomeric and subtelomeric rearrangements observed previously in telomerase-defective survivor strains⁸ (data not shown).

Mutations in *MLH1* or *PMS1* also significantly enhanced the growth of telomerase-defective strains (see Supplementary Information). Loss of either *MSH3* or *MSH6* alone had little or no effect, but a *msh3-Δ msh6-Δ* double mutation enhanced survival (see Supplementary Information), consistent with the previously reported partly redundant roles of these genes in mismatch recognition. Thus, the enhancement of telomerase-independent survival is caused by loss of the MMR pathway, rather than an effect specific to *MSH2*. Msh2 and Msh3 proteins also participate, with the Rad1–Rad10 endonuclease complex, in direct repeat recombination and gene-conversion events, by removing non-homology from DNA ends during recombination¹⁸. *rad1-Δ* and *rad10-Δ* mutations had no effect on the frequency of telomerase-independent survival (Fig. 2a, b), however, consistent with observations showing that

these gene products have little or no anti-recombination activity¹⁹.

Because defects in MMR increase mutation rates, we also considered the possibility that the increase in telomerase-independent survival might be the secondary consequence of mutations in gene(s) that influence recombination at telomeres. A *pol3-01* mutation, which renders DNA polymerase- δ defective in proof-reading and generates mutation rates similar to those of a *msh2- Δ* strain²⁰ without rendering cells defective in MMR²¹, had no effect on the growth of a telomerase-defective strain (Fig. 2c). Furthermore, an *est2- Δ* *msh2- Δ* strain grown for ~50 generations re-acquired the growth characteristics of an *est2- Δ* strain if the wild-type *MSH2* gene was re-introduced (data not shown). These results, combined with the competitive advantage exhibited by *est2- Δ* *msh2- Δ* strains early in their senescence progression (Fig. 1e), indicate that the growth advantage observed in the absence of *MSH2* is not the result of suppressor mutation(s).

Finally, *RAD52*-dependent recombination has been shown to be required for telomerase-independent growth^{8,9,11}; as expected, the enhancement of telomerase-independent survival by a MMR defect also requires *RAD52* (data not shown). These results support the hypothesis that removing the block to homeologous recombination enhances telomerase-independent survival.

Mutations in *hMSH2* and *hMLH1*, and to a lesser extent *hPMS1*, have been found in the familial cancers of HNPCC patients, as well as sporadic tumours of the colon, breast and other tissues³⁻⁵. To determine whether the results that we observe in the budding yeast *S. cerevisiae* could be relevant to human tumorigenesis, we examined the effects of two yeast *MSH2* missense mutations (*msh2*_{G693D} and *msh2*_{S656P}) that are identical to mutations identified in HNPCC *hms2* kindreds¹³. These *msh2* mutations, when introduced into a telomerase-defective strain, enhanced survival to a degree comparable to the enhancement observed when *MSH2* was deleted (Fig. 3). Therefore, a specific mutational event corresponding to that

observed in a subset of HNPCC tumours can enhance the growth of telomerase-defective yeast cells.

Because *S. cerevisiae* telomeres are composed of irregular G₁₋₃T repeats²², there is a high potential for mismatches when terminal sequences recombine. In contrast, human telomerase adds perfect T₂AG₃ repeats onto chromosome termini, although the proximal region of human telomeres is composed of variant repeats^{14,15}. To

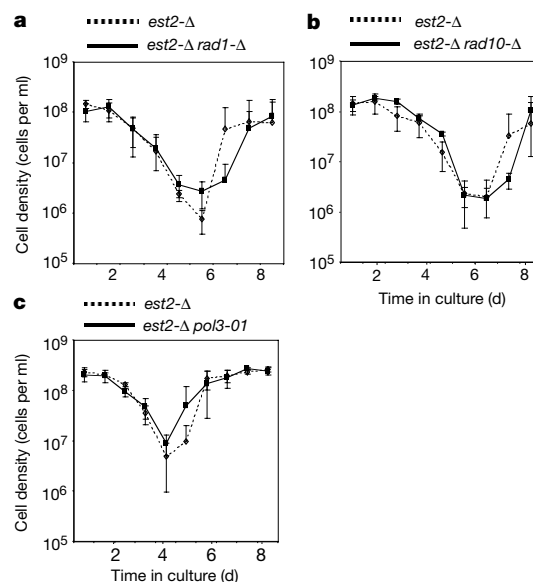


Figure 2 *rad1- Δ* , *rad10- Δ* or *pol3-01* have no significant effect on *est2- Δ* viability, as measured by liquid growth assays. **a**, *rad1- Δ* ; **b**, *rad10- Δ* ; **c**, *pol3-01*. Results are from three samples of each indicated genotype.

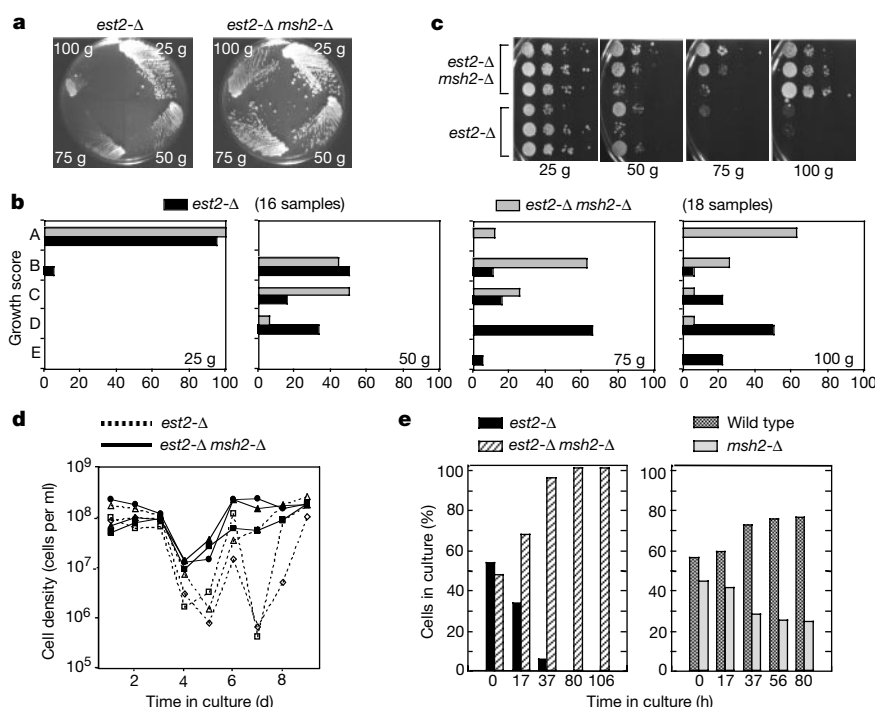


Figure 1 Loss of *MSH2* function enhances the growth of a telomerase-defective strain. **a**, Serial single-colony streak-outs for *est2- Δ* and *est2- Δ msh2- Δ* strains after ~25, 50, 75 and 100 generations (g) of growth. **b**, Distribution of growth scores, shown as percentages, for 16 *est2- Δ* and 18 *est2- Δ msh2- Δ* strain streak-outs after ~25, 50, 75, and 100 generations. **c**, Tenfold dilution series of *est2- Δ* and *est2- Δ msh2- Δ* strains after ~25, 50, 75 and 100 generations. **d**, Liquid growth assays for three independent

samples each of *est2- Δ* and *est2- Δ msh2- Δ* ($P \leq 0.01$ for days 4 and 5). **e**, Liquid competition experiments between *est2- Δ* and *est2- Δ msh2- Δ* (left) or wild-type and *msh2- Δ* (right) strains; the genotype of each pair of strains was identical except for the *msh2- Δ* ::*Kan^r* disruption, which does not confer a growth advantage under these experimental conditions.

determine whether MMR-dependent repression of the survival of telomerase-defective strains is relevant in other species, we examined *Kluyveromyces lactis*, another budding yeast that undergoes RAD52-dependent survival after the loss of telomerase⁹. *K. lactis* telomerase adds perfect G-rich repeats onto a telomere tract that is comparable in length to that of *S. cerevisiae*²³. However, the sequence of *K. lactis* chromosome termini is degenerate in the subtelomeric regions, and thus could be a likely target of the anti-recombination activity of the MMR machinery.

To examine the effects of the loss of MMR on telomerase-independent survival in *K. lactis*, we cloned the *kIMSH2* gene. The *K. lactis* Msh2 protein showed 76% similarity and 59% identity to *S. cerevisiae* Msh2, and disruption of *kIMSH2* produced a ~100-

fold increase in mutation rate (data not shown). When a *ter1-Δ* strain (disrupted for the RNA subunit of *K. lactis* telomerase) was compared with a *ter1-Δ msh2-Δ* strain, the *ter1-Δ msh2-Δ* strain exhibited a clear growth advantage after ~80 generations of growth (Fig. 4a; and data not shown). A growth advantage was also revealed early during senescence by a competition test: in five independent experiments that monitored continuous log-phase growth of a mixture of *ter1-Δ* and *ter1-Δ msh2-Δ* strains, the *ter1-Δ msh2-Δ* strain overgrew the culture in each case (Fig. 4b). Therefore, even in an organism in which the telomeric repeat region is not completely degenerate, elimination of MMR enhances telomerase-independent proliferation.

The barrier to homeologous recombination has been proposed

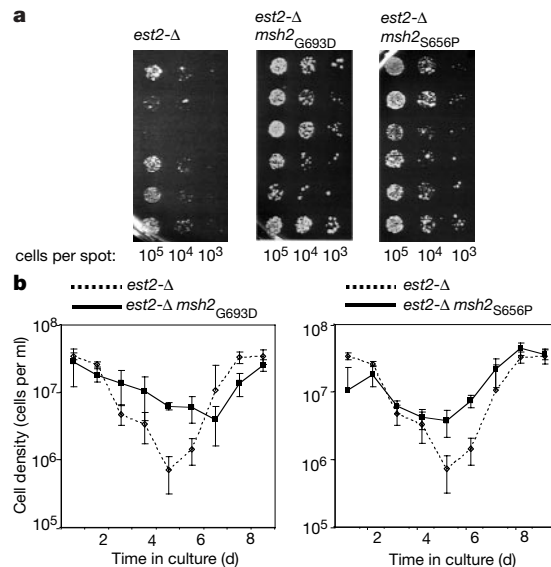


Figure 3 HNPCC-like mutations in *S. cerevisiae* *MSH2* enhance *est2-Δ* viability. **a**, Growth of tenfold dilution series for six samples of each indicated genotype after ~75 generations of growth. **b**, Liquid growth assays for three independent samples of each indicated

genotype. $P \leq 0.07$, 0.002 and 0.02 for days 4–6, respectively, for *est2-Δ* versus *est2-Δ msh2^{G693D}*; and $P \leq 0.06$ and 0.002 for days 5 and 6, respectively, for *est2-Δ* versus *est2-Δ msh2^{S656P}*.

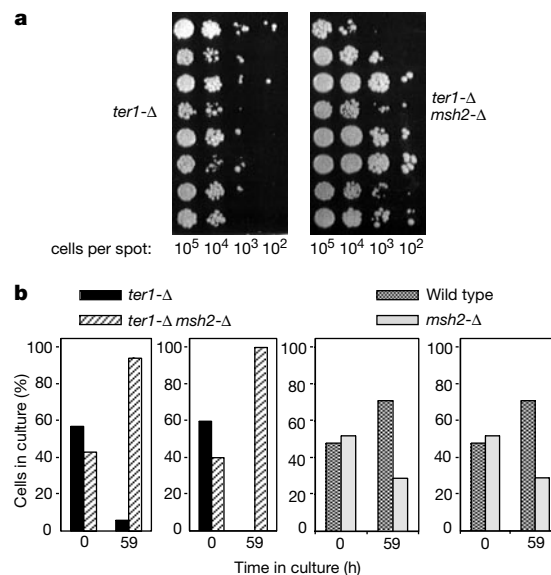


Figure 4 Loss of mismatch repair in *K. lactis* enhances *ter1-Δ* viability. **a**, Tenfold dilution series of *ter1-Δ* and *ter1-Δ msh2-Δ* strains after ~80 generations of growth. **b**, Representative liquid competition experiments between *ter1-Δ* and *ter1-Δ msh2-Δ*

(left) or wild-type and *msh2-Δ* (right) strains; the genotype of each pair of strains was identical, except for the presence or absence of the *MSH2* gene.

to preserve chromosomal integrity by preventing undesirable exchanges between chromosomes². Our results show that removal of this barrier can, under certain conditions, promote cell proliferation. A similar enhancement may contribute to cell immortalization in the absence of telomerase in MMR-defective human cells, either by prolonging the critical period during which telomerase reactivation occurs or potentially alleviating the requirement for telomerase reactivation altogether. Such a possibility might have implications for potential anti-telomerase therapeutics for some classes of tumours. □

Methods

Strains

Details on the construction of *S. cerevisiae* and *K. lactis* strains are given in the Supplementary Information.

Serial colony streak-out analysis

We assessed growth phenotype by four successive single-colony streak-outs, starting with either freshly dissected spores of the appropriate genotype for *S. cerevisiae*, or freshly generated *ter1* isolates for *K. lactis*. Streak-outs were stored at room temperature during the course of the experiment, and successive streak-outs were reassembled on the same plate without notation of the genotype and scored after 2 days of growth at 30 °C. For *S. cerevisiae*, we used an arbitrary scale of A to E: A and E corresponded to 'wild-type growth' and 'no growth', respectively, whereas intermediate ratings (B, C and D) depended on both relative colony size and degree of heterogeneity—properties that have been noted previously to correlate with the severity of the growth phenotype of a telomerase-defective strain^{8,16}. For serial dilution assays, colonies from successive streak-outs were resuspended in water and diluted to 10⁷ per ml, as determined by haemocytometer cell counts. Tenfold serial dilutions were assembled in microtitre dishes, stamped onto rich media and evaluated for growth after a 2-day incubation at 30 °C.

Competition experiments

We mixed 10⁴ cells per ml from two different log-phase cultures from freshly dissected colonies (that is, early in their senescence progression) together in rich media and re-grew them to mid-log (~21 h for early time points, longer for late senescing cultures). Cell density was determined by haemocytometer, plated for subsequent genotype determination, and then diluted to 2 × 10⁴ cells per ml and re-grown. For *S. cerevisiae* experiments, the genotype of the culture of ~1,000 colonies was assessed by replica-plating onto plates containing the drug G418, whereas the genotype of ~50 colonies was determined by colony polymerase chain reaction for the *K. lactis* experiments.

Liquid growth assays

We used a freshly dissected colony to inoculate a culture at 10⁴ cells per ml and grew them overnight to mid-log phase. Cell counts were determined by haemocytometer and the culture was re-diluted to 10⁵ cells per ml, and re-grown for ~21 h.

Received 29 January 2000; accepted 26 February 2001.

- Modrich, P. & Lahue, R. S. Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem.* **65**, 101–133 (1996).
- Rayssiguier, C., Thaler, D. S. & Radman, M. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342**, 389–401 (1989).
- Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215–1225 (1993).
- Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
- Modrich, P. Mismatch repair, genetic stability and cancer. *Science* **266**, 1959–1960 (1994).
- Hahn, W. C. *et al.* Inhibition of telomerase limits the growth of human cancer cells. *Nature Med.* **5**, 1164–1170 (1999).
- Shay, J. W. & Bacchetti, S. A survey of telomerase activity in human cancer. *Eur. J. Cancer* **33**, 787–791 (1997).
- Lundblad, V. & Blackburn, E. H. An alternate pathway for yeast telomere maintenance rescues *est1* senescence. *Cell* **73**, 347–360 (1993).
- McEachern, M. J. & Blackburn, E. H. Cap-prevented recombination between terminal telomeric repeat arrays (telomere CPR) maintains telomeres in *Kluyveromyces fragilis* lacking telomerase. *Genes Dev.* **10**, 1822–1834 (1996).
- Nakamura, T. M., Cooper, J. P. & Cech, T. R. Two modes of survival of fission yeast without telomerase. *Science* **282**, 493–496 (1998).
- Teng, S. C. & Zakian, V. A. Telomere–telomere recombination is an efficient bypass pathway for telomere maintenance in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19**, 8083–8093 (1999).
- Dunham, M. A., Neumann, A. A., Fasching, C. L. & Reddel, R. R. Telomere maintenance by recombination in human cells. *Nature Genet.* **26**, 447–450 (2000).
- Studamire, B., Price, G., Sugawara, N., Haber, J. E. & Alani, E. Separation-of-function mutations in *Saccharomyces cerevisiae* MSH2 that confer mismatch repair defects but do not affect nonhomologous-tail removal during recombination. *Mol. Cell. Biol.* **19**, 7558–7567 (1999).
- Flint, J. *et al.* Sequence comparisons of human and yeast telomeres identified structurally distinct subtelomeric domains. *Human Mol. Genet.* **6**, 1305–1313 (1997).
- Allshire, R. C., Dempster, M. & Hastie, N. D. Human telomeres contain at least three types of G-rich repeat distributed non-randomly. *Nucleic Acids Res.* **17**, 4611–4627 (1989).

- Lundblad, V. & Szostak, J. W. A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell* **57**, 633–643 (1989).
- Lingner, J. *et al.* Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276**, 561–567 (1997).
- Fishman-Lobell, J. & Haber, J. E. Removal of nonhomologous DNA ends in double-strand break recombination: the role of the yeast ultraviolet repair gene *RAD1*. *Science* **258**, 480–484 (1992).
- Nicholson, A., Hendrix, M., Jinks-Robertson, S. & Crouse, G. F. Regulation of mitotic homeologous recombination in yeast. Function of mismatch repair and nucleotide excision repair genes. *Genetics* **154**, 133–146 (2000).
- Tran, H. T., Gordenin, D. A. & Resnick, M. A. The 3'→5' exonucleases of DNA polymerases delta and epsilon and the 5'→3' exonuclease Exo1 have major roles in postreplication mutation avoidance in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19**, 2000–2007 (1999).
- Datta, A. *et al.* Checkpoint-dependent activation of mutagenic repair in *Saccharomyces cerevisiae* *pol3-01* mutants. *Mol. Cell* **6**, 593–603 (2000).
- Shampay, J., Szostak, J. W. & Blackburn, E. H. DNA sequences of telomeres maintained in yeast. *Nature* **310**, 154–157 (1984).
- McEachern, M. J. & Blackburn, E. H. A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts. *Proc. Natl Acad. Sci. USA* **91**, 3453–3457 (1994).

Supplementary information is available on Nature's World Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. Plon for the initial suggestion to test the effects of a MMR defect in telomerase-defective strains; M. McEachern and E. Alani for the gifts of strains and plasmids; and J. Haber, S. Rosenberg, P. Hastings and S. Evans for critical comments. This work was supported by a US Army Breast Cancer Research Predoctoral Fellowship to A.R. and grants from the NIH and Houston Cancer Fighters to V.L.

Correspondence and requests for materials should be addressed to V.L. (e-mail: lundblad@bcm.tmc.edu). The *K. lactis* MSH2 sequence has been deposited in GenBank under accession code AF332582.

A freely diffusible form of Sonic hedgehog mediates long-range signalling

Xin Zeng*, John A. Goetz*, Liza M. Suber*, William J. Scott Jr†, Claire M. Schreiner† & David J. Robbins*

*Department of Molecular Genetics, Biochemistry and Microbiology, University of Cincinnati College of Medicine, 231 Albert Sabin Way, Cincinnati, Ohio 45267, USA

†Children's Hospital Research Foundation, Division of Developmental Biology, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA

The secreted protein Sonic hedgehog (Shh) exerts many of its patterning effects through a combination of short- and long-range signalling^{1–3}. Three distinct mechanisms, which are not necessarily mutually exclusive, have been proposed to account for the long-range effects of Shh: simple diffusion of Shh, a relay mechanism in which Shh activates secondary signals, and direct delivery of Shh through cytoplasmic extensions, termed cytonemes. Although there is much data (using soluble recombinant Shh (ShhN)) to support the simple diffusion model of long-range Shh signalling^{1,2}, there has been little evidence to date for a native form of Shh that is freely diffusible and not membrane-associated. Here we provide evidence for a freely diffusible form of Shh (s-ShhNp) that is cholesterol modified, multimeric and biologically potent. We further demonstrate that the availability of s-ShhNp is regulated by two functional antagonists of the Shh pathway, Patched (Ptc) and Hedgehog-interacting protein (Hip)^{4–6}. Finally, we show a gradient of s-ShhNp across the anterior–posterior axis of the chick limb, demonstrating the physiological relevance of s-ShhNp.

We thought that Shh was capable of directly initiating long-range signalling; however, this had been undetected owing to the lack of

high-affinity antibodies. We therefore began our search for active soluble Shh by taking advantage of a sensitive cell-based bioassay for Shh, the differentiation of C3H10T1/2 cells⁷. For an initial source of soluble Shh we transfected Bosc 23 cells with constructs encoding full-length Shh, ShhN, or a control vector. The conditioned media from the various transfected Bosc cells was added to C3H10T1/2 cells to measure Shh biological activity (Fig. 1a). The conditioned media from Bosc cells transfected with full-length Shh differentiated the C3H10T1/2 cells and had activity comparable to media from ShhN-transfected cells. As previously reported^{8–10}, we found that most of the processed lipid-modified form of Shh (ShhNp) was cell-associated whereas ShhN, which is not lipid modified, was mostly soluble. To determine the amount of soluble Shh we carried out an immunoprecipitation, followed by immunoblotting with Shh antibodies (Fig. 1b). We detected a small amount of soluble Shh from full-length Shh-conditioned media and a large amount of ShhN from ShhN-conditioned media. Soluble Shh migrated at the same apparent molecular weight (on a SDS–polyacrylamide gel) as ShhNp from lysates of transfected Bosc cells (Fig. 1b) and faster than ShhN, consistent with it being modified by cholesterol. For this reason, and others (see Supplementary Information), we refer to this form of soluble Shh as s-ShhNp, to distinguish it from the membrane-associated form of ShhNp.

We subjected the Shh-conditioned media to ultracentrifugation, to rule out the possibility that s-ShhNp was only Shh-bound to membranous vesicles. We divided the s-ShhNp into two aliquots. One aliquot was set aside, while the other was centrifuged at 100,000g for 120 min. Any resulting pellet was resuspended with an equal volume of media. There was no change in alkaline phosphatase (AP) activity between the centrifuged and non-centrifuged conditioned media (Fig. 1c), suggesting that the Shh in the media is indeed soluble.

To verify that the biological activity in the Shh-conditioned media is due to Shh we used three distinct assays of Shh activity: activation of AP in C3H10T1/2 cells; activation of *Ptc* transcription in C3H10T1/2 cells; and activation of a Gli-luciferase reporter in Shh-Light2 cells¹¹ (see Methods). These assays were performed in the presence or absence of two specific Shh inhibitors—the Shh neutralizing monoclonal antibody 5E1 (ref. 12) and the teratogen

cyclopamine^{13,14}. In the AP assay, aliquots of conditioned media were pre-incubated with the 5E1 monoclonal antibody or an isotype-matched control, then added to C3H10T1/2 cells, followed by AP staining (Fig. 2a). The biological activity of the Shh-conditioned media was almost completely inhibited by the 5E1 monoclonal antibody, but unaffected by immunoglobulin- γ (IgG). We also examined the ability of s-ShhNp to activate *Ptc* transcription in C3H10T1/2 cells, as *Ptc*, unlike AP, is thought to be a direct target gene of Shh. s-ShhNp was able to activate *Ptc* transcription in a manner comparable to ShhN and this activation was inhibited by cyclopamine (Fig. 2b), as was induction of AP (see Supplementary Information). Furthermore, s-ShhNp activated Shh-Light2 cells approximately tenfold more than control conditioned media, and most of this activity was specifically inhibited by the 5E1 monoclonal antibody (see Supplementary Information). Thus, the biological activity of the Shh-conditioned media is due to s-ShhNp, as two mechanistically distinct antagonists of the Shh pathway can inhibit the activity of the conditioned media in both direct and indirect assays of Shh activity.

It has been thought that the Shh receptor *Ptc* and *Hip* are functional antagonists of Shh family members^{4–6}. Although it has been suggested that *Ptc* and *Hip* antagonize Shh signalling by sequestering Shh, this sequestration has never been demonstrated. To examine the effects of these Shh antagonists on the biological activity of s-ShhNp we co-transfected cells with Shh and *Ptc*, *Hip* or

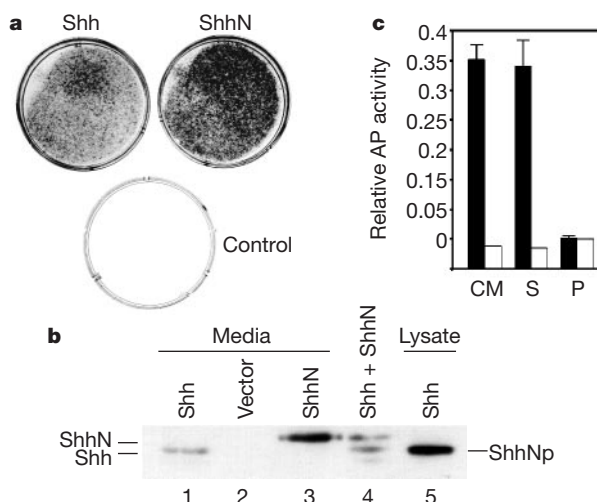


Figure 1 Identification of a soluble form of Shh. **a**, Conditioned media from Shh, ShhN or vector-transfected cells added to C3H10T1/2 cells and stained for AP activity. **b**, Shh in conditioned media from full-length Shh-transfected cells detected (lane 1) by immunoprecipitation followed by western blot analysis. Lane 4 is a mixture of the samples loaded in lanes 1 and 3. **c**, Conditioned media from full-length Shh-transfected cells induced similar AP activity before (CM) and after (S) ultracentrifugation (100,000g, 120 min). Black bars, Shh-conditioned media; white bars, control media.

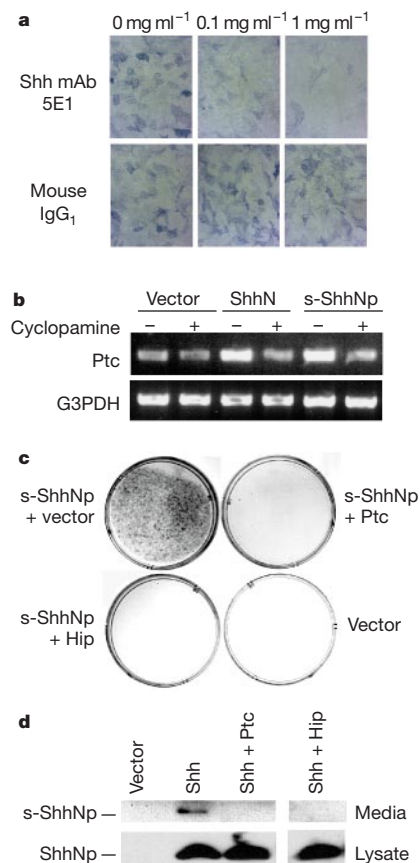


Figure 2 Soluble Shh activity is inhibited by Shh functional antagonists. **a**, Conditioned media from Shh-transfected cells incubated with dilutions of 5E1 monoclonal antibody (mAb) or IgG1 added to C3H10T1/2 cells and stained for AP induction. **b**, C3H10T1/2 cells incubated with indicated conditioned media with or without cyclopamine (2.4 μ M), followed by RT–PCR analysis. **c**, Conditioned media from 293T cells co-transfected with the indicated plasmids and assayed for induction of AP activity. **d**, Aliquots of the conditioned media in **c** analysed for the presence of s-ShhNp. Note the presence of equal amounts of ShhNp in the corresponding cell lysates.

a vector control, and then tested the conditioned media for biological activity. Conditioned media from Shh-transfected cells produced in the presence of Ptc or Hip has reduced biological activity compared with media from Shh co-transfected with a vector control (Fig. 2c). This decrease in biological activity correlates with decreased s-ShhNp in the conditioned media from Ptc- or Hip-expressing cells (Fig. 2d). Thus Ptc and Hip, two physiologically relevant inhibitors of the Shh pathway, are able to reduce the activity of s-ShhNp in the conditioned media by reducing its availability.

One reason why it has been difficult to detect Shh far from its site of synthesis may be that s-ShhNp is very potent, requiring Shh concentrations below levels detectable with current antibody reagents. Most of the data concerning the potency of Shh has been generated using recombinant ShhN, which is an order of magnitude less potent than purified ShhNp⁸. To compare the potency of ShhN to s-ShhNp we transfected Bosc cells with Shh, ShhN or a vector control, then collected the various conditioned media. We found that a 50% dilution of the ShhN-conditioned media had comparable biological activity to full-strength s-ShhNp-conditioned media (Fig. 3a). However, comparable amounts of s-ShhNp and ShhN immunoreactivity were only seen when ShhN media was diluted thirty times (Fig. 3b), indicating that s-ShhNp is greater than fifteen times more potent than ShhN.

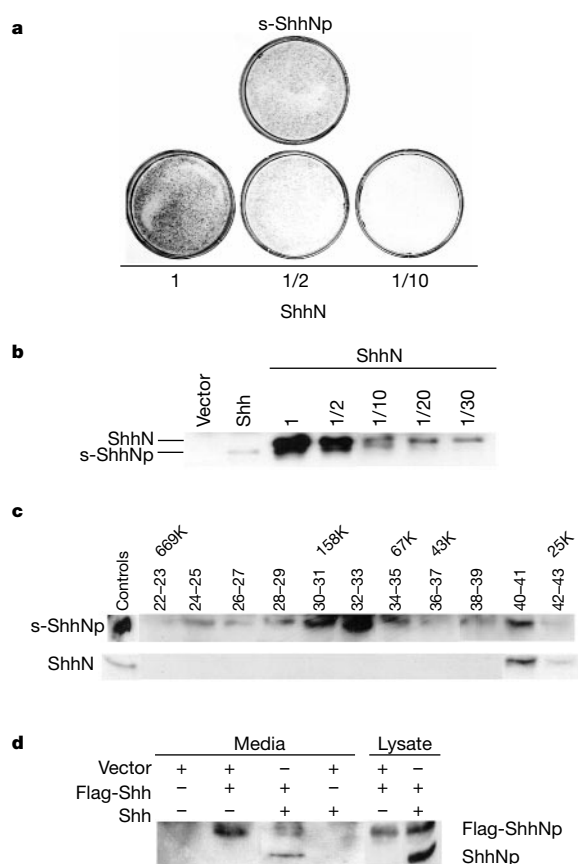


Figure 3 s-ShhNp is a highly potent, multimeric form of ShhNp. **a**, **b**, s-ShhNp or indicated dilutions of ShhN-conditioned media used to determine AP activity (**a**) or the relative amount of Shh present (**b**) in various conditioned media. **c**, ShhN- or s-ShhNp-conditioned media fractionated by gel filtration. The indicated fractions were pooled and analysed for the presence of Shh. The fractions in which molecular weight standards elute are indicated as relative molecular mass (*M*) values. **d**, Conditioned media from cells transfected with the indicated plasmids, collected and immunoprecipitated with the anti-Flag M2 monoclonal antibody, and then immunoblotted using the 5E1 monoclonal antibody.

To address how s-ShhNp, a protein with two lipophilic modifications (see Supplementary Information), is soluble in an aqueous environment we proposed that Shh might bind additional proteins to bury its hydrophobic moieties. We therefore isolated conditioned media from cells transfected with full-length Shh or ShhN and subjected it to analysis by gel filtration chromatography. These conditioned media were isolated under serum-free conditions in which they are still biologically active to verify that s-ShhNp was not just binding to proteins present in serum. s-ShhNp migrated at about six times its native molecular weight (Fig. 3c). In contrast, ShhN migrated through this column close to its predicted molecular weight. To determine whether Shh could multimerize with itself to form this large Shh complex, we immunoprecipitated s-ShhNp from the conditioned media of cells co-transfected with Shh and a Flag-tagged Shh construct. Antibodies to the Flag epitope were able to co-immunoprecipitate untagged Shh, suggesting that Shh multimerizes with itself (Fig. 3d).

To verify that s-ShhNp could be isolated from a physiologically relevant source of Shh we looked at the chick limb-bud system^{15–19}. SHH expression co-localizes with polarizing activity at various stages of limb development, where it is thought to be responsible for the polarizing activity associated with this posterior region of the limb bud^{20,21}. Therefore, similar-sized fragments of posterior limb-bud tissue were dissected out and cultured in DMEM to allow s-ShhNp to diffuse out. Conditioned media from posterior regions of the limb bud strongly induced AP activity, and this biological activity was blocked by the 5E1 monoclonal antibody and cyclopamine (see Supplementary Information). Conditioned media from posterior tissue was also able to activate Shh-Light2 cells 3–4 times over that of control media (Fig. 4a, top). Additionally, s-ShhNp could be detected in immunoprecipitates from posterior tissue

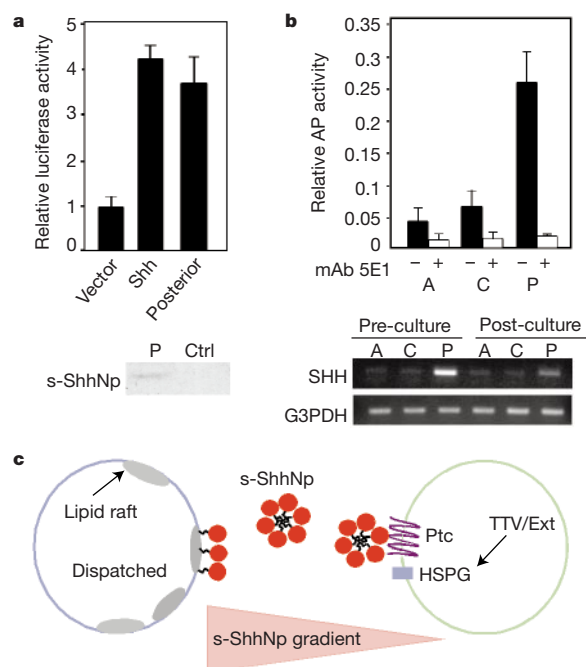


Figure 4 s-ShhNp exists in an anterior-posterior gradient. **a**, Conditioned media from chick limb-bud posterior tissue analysed for Gli-luciferase activity (top) or the presence of s-ShhNp (bottom). s-ShhNp was immunoprecipitated from posterior tissue (P) conditioned media or control (Ctrl) media lacking tissue. **b**, Conditioned media from the anterior (A), central (C) or posterior (P) third of limb buds added to C3H10T1/2 cells in the presence (white bars) or absence (black bars) of 5E1 monoclonal antibody (mAb). The insert shows an RT-PCR analysis of SHH mRNA levels under the indicated conditions. **c**, Model of s-ShhNp activity in long-range signalling. HSPG, heparin sulphate proteoglycan.

conditioned media, but was absent from media conditioned in the absence of limb tissue (Fig. 4a, bottom). This endogenous form of soluble Shh migrated, on SDS-PAGE, in a manner consistent with it also being cholesterol modified (data not shown). These results demonstrate that s-ShhNp can also be obtained from a physiologically relevant source of Shh, the posterior region of chick limb buds. Although we suggest that this endogenous form of soluble Shh is the same as that characterized in Figs 1–3, we cannot rule out that it might be different.

Given the potency of s-ShhNp it may previously have been difficult to detect a SHH gradient across the limb bud using immunohistochemistry. Therefore, we decided to re-investigate the presence of a soluble Shh gradient across the limb bud using our Shh bioassay. Similar-sized pieces of posterior, central or anterior tissue were incubated in tissue culture media to isolate s-ShhNp (Fig. 4b). Under conditions in which our AP assay is linear (data not shown) strong Shh activity was obtained from posterior tissues, approximately 5–6 times less activity was obtained from central tissues, and even less Shh activity was obtained from anterior regions of the limb bud. Most of the s-ShhNp activity obtained from all regions of the limb bud was blocked by the Shh inhibitory monoclonal antibody. By using polymerase chain reaction after reverse transcription of RNA (RT-PCR) we showed that *SHH* transcription is not upregulated in limb tissue after culture. This rules out the possibility that the Shh activity seen in the anterior and central fragments of limb tissue resulted from the upregulation of endogenous Shh messenger RNA. Thus, s-ShhNp is produced from a physiologically relevant source of Shh, the limb bud, where it forms a gradient from posterior regions across the midline of the limb bud and perhaps even into more anterior regions of the limb bud. Although we interpret our data as showing a gradient of s-ShhNp across the limb bud (see also Supplementary Information), we have not yet ruled out the possibility of additional factors inducing low levels of Shh expression in our reporter cell line.

We provide evidence for a physiologically relevant soluble form of Shh, s-ShhNp, which meets many of the criteria necessary for a long-range signalling molecule²². s-ShhNp is a new form of Shh in that it is a freely diffusible multimeric form of ShhNp. We suggest a model in which s-ShhNp diffuses from its site of synthesis to provide long-range patterning information (Fig. 4c), such as seen in the limb bud or in motor neuron differentiation^{1,2}. s-ShhNp would be formed by concentrating ShhNp at lipid rafts, where it multimerizes with its lipid attachments sequestered in the interior of the multimer. This soluble multimer of ShhNp would then diffuse far from its site of synthesis to engage directly in long-range signalling. We further speculate that mammalian homologues of the *Drosophila* protein Dispatched (DISP)²³ might be involved in either the packaging of s-ShhNp or the targeting of ShhNp to lipid rafts, given that amorphic *disp* mutants are insensitive to HhNp. Additionally, Tout velu (TTV), an enzyme involved in heparin sulphate biosynthesis, has been implicated as a necessary component in Hh-receiving cells²⁴. Although we have not examined the function that mammalian TTV homologues have in Shh target cells, we do present biochemical data consistent with the proposed function of TTV in *Drosophila*. That is, *ttv* mutations are insensitive to HhNp but responsive to HhN, suggesting that Hh's lipid modifications are necessary for TTV to exert its normal biological effects. We present data showing a central difference between ShhN and ShhNp: ShhNp can form large, stable multimers whereas ShhN cannot. Therefore, TTV homologues might regulate the biosynthesis of a molecule that is necessary to recognize s-ShhNp but is not necessary for signalling by ShhN. □

Methods

Constructs

The ShhN construct encoding the amino-terminal fragment (amino acids 1–201) of rat Shh was generated by deleting an internal 3' fragment from the pMT-Shh expression

vector (a gift from T. M. Jessell). The 2× Flag-tagged Shh construct was provided by R. K. Guy²⁵. The Myc-tagged mouse Ptc complementary DNA expression vector was provided by R. Johnson. The Myc-tagged Hip cDNA expression vector was provided by A. McMahon. Cyclopamine was provided by W. Gaffield.

Transfection and production of conditioned media

Bosc 23 or 293T cells were maintained in DMEM with 10% fetal bovine serum. Shh, ShhN, or a control expression vector were transfected into cells using lipofectamine (Gibco). Conditioned media was collected 48 h after transfection and filtered through a 0.2-μm filter before any analysis. The cells were lysed in Nonidet P-40 buffer (150 mM NaCl, 50 mM HEPES pH 7.4, 1.5 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol, 1% Nonidet P-40, 1:1,000 dilution of a protease inhibitor mixture (PI)). PI contains 1 mM benzamide, 1 mg ml⁻¹ aprotinin, 1 mg ml⁻¹ leupeptin and 1 mg ml⁻¹ pepstatin A in 100% ethanol. The co-transfection of Shh with Ptc or Hip cDNA was carried out in 293T cells. The Shh-Light2 cells¹¹ (an NIH-3T3-derived cell line containing an integrated Gli-luciferase reporter gene) was provided by P. A. Beachy. The luciferase activity was normalized to a Renilla control and the results are expressed as mean ± s.d. for two experiments. Metabolic labelling of 293T cells with [³H]-cholesterol was done essentially as described²⁶. The [³H]-labelled conditioned media and the cell lysates were collected 48 h after transfection and immunoprecipitated with 5E1. We separated the immunoprecipitates by SDS-PAGE for analysis as described²⁷.

Gel filtration chromatography

Serum-free conditioned media in Optimen (Gibco) from 293T cells transfected with Shh or ShhN were collected and ultracentrifuged at 100,000g for 2 h. The supernatants were concentrated using a Centrprep YM-10 (Amicon) and loaded on to a Superose 12 (Pharmacia) gel filtration column that had been equilibrated with PBS/0.001% NP-40. Every two fractions were pooled and immunoprecipitated with the 5E1 monoclonal antibody, resolved by SDS-PAGE, and immunoblotted with the 5E1 monoclonal antibody.

Dissection and culture of chick limb buds

Regions of limb buds were dissected from Hamburger and Hamilton stage 22 chick embryos as described²⁸ and cultured in DMEM with 10% FBS for 16 h. The media was collected, centrifuged at 100,000g for 2 h and added (100 μl well⁻¹) to C3H10T1/2 cells cultured in a 96-well plate for 5 d before measuring AP activity.

C3H10T1/2 alkaline phosphatase (AP) assay

Histochemical staining for AP was performed as previously described⁷. The liquid assay for AP activity was performed by measuring absorbance at 405 nm using the chromogenic substrate *p*-nitrophenyl phosphate (Sigma)⁷. The results are expressed as mean ± s.d. (*n* ≥ 3).

Received 7 March; accepted 16 May 2001.

- Johnson, R. L. & Tabin, C. The long and short of hedgehog signaling. *Cell* **81**, 313–316 (1995).
- McMahon, A. P. More surprises in the Hedgehog signaling pathway. *Cell* **100**, 185–188 (2000).
- Chuang, P. & Kornberg, T. B. On the range of hedgehog signaling. *Curr. Opin. Genet. Dev.* **10**, 515–522 (2000).
- Chen, Y. & Struhl, G. Dual roles for patched in sequestering and transducing Hedgehog. *Cell* **87**, 553–563 (1996).
- Chuang, P. T. & McMahon, A. P. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature* **397**, 617–621 (1999).
- Goodrich, L. V., Jung, D., Higgins, K. M. & Scott, M. P. Overexpression of ptc1 inhibits induction of Shh target genes and prevents normal patterning in the neural tube. *Dev. Biol.* **211**, 323–334 (1999).
- Nakamura, T. *et al.* Induction of osteogenic differentiation by hedgehog proteins. *Biochem. Biophys. Res. Commun.* **237**, 465–469 (1997).
- Pepinsky, R. B. *et al.* Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273**, 14037–14045 (1998).
- Bumcrot, D. A., Takada, R. & McMahon, A. P. Proteolytic processing yields two secreted forms of sonic hedgehog. *Mol. Cell. Biol.* **15**, 2294–2303 (1995).
- Porter, J. A. *et al.* The product of hedgehog autoproteolytic cleavage active in local and long-range signalling. *Nature* **374**, 363–366 (1995).
- Taipale, J. *et al.* Effects of oncogenic mutations in Smoothened and Patched can be reversed by cyclopamine. *Nature* **406**, 1005–1009 (2000).
- Ericson, J., Morton, S., Kawakami, A., Roelink, H. & Jessell, T. M. Two critical periods of Sonic Hedgehog signaling required for the specification of motor neuron identity. *Cell* **87**, 661–673 (1996).
- Cooper, M. K., Porter, J. A., Young, K. E. & Beachy, P. A. Teratogen-mediated inhibition of target tissue response to Shh signaling. *Science* **280**, 1603–1607 (1998).
- Incardona, J. P., Gaffield, W., Kapur, R. P. & Roelink, H. The teratogenic Veratrum alkaloid cyclopamine inhibits Sonic hedgehog signal transduction. *Development* **125**, 3553–3562 (1998).
- Tickle, C. Morphogen gradients in vertebrate limb development. *Semin. Cell Dev. Biol.* **10**, 345–351 (1999).
- Yang, Y. *et al.* Relationship between dose, distance and time in Sonic Hedgehog-mediated regulation of anteroposterior polarity in the chick limb. *Development* **124**, 4393–4404 (1997).
- Lopez-Martinez, A. *et al.* Limb-patterning activity and restricted posterior localization of the amino-terminal product of Sonic hedgehog cleavage. *Curr. Biol.* **5**, 791–796 (1995).
- Marti, E., Takada, R., Bumcrot, D. A., Sasaki, H. & McMahon, A. P. Distribution of Sonic hedgehog peptides in the developing chick and mouse embryo. *Development* **121**, 2537–2547 (1995).
- Wang, B., Fallon, J. F. & Beachy, P. A. Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell* **100**, 423–434 (2000).
- Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
- Chang, D. T. *et al.* Products, genetic linkage and limb patterning activity of a murine hedgehog gene. *Development* **120**, 3339–3353 (1994).

22. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25**, 1–47 (1969).
23. Burke, R. *et al.* Dispatched, a novel sterol-sensing domain protein dedicated to the release of cholesterol-modified hedgehog from signaling cells. *Cell* **99**, 803–815 (1999).
24. The, I., Bellaiche, Y. & Perrimon, N. Hedgehog movement is regulated through tout velu-dependent synthesis of a heparan sulfate proteoglycan. *Mol. Cell* **4**, 633–639 (1999).
25. Guy, R. K. Inhibition of sonic hedgehog autotransformation in cultured mammalian cells by sterol deprivation. *Proc. Natl Acad. Sci. USA* **97**, 7307–7312 (2000).
26. Porter, J. A., Young, K. E. & Beachy, P. A. Cholesterol modification of hedgehog signaling proteins in animal development. *Science* **274**, 255–259 (1996).
27. Robbins, D. J. *et al.* Hedgehog elicits signal transduction by means of a large complex containing the kinesin-related protein COSTAL2. *Cell* **90**, 225–234 (1997).
28. Bell, S. M., Schreiner, C. M. & Scott, W. J. Transspecies grafting as a tool to understand the basis of murine developmental limb abnormalities. *Methods Mol. Biol.* **136**, 219–226 (2000).

addendum

A universal scaling law for atomic diffusion in condensed matter

M. Dzugutov

Nature **381**, 137–139 (1996).

The author inadvertently omitted to mention an earlier result¹ relating diffusion D , excess entropy s , density ρ and temperature T in liquids as $D\rho^{1/3}(m/k_B T)^{1/2} = Ae^{Bs}$. It was interpreted² as a convenient approximation of an essentially algebraic relation with B changing from 0.65 for hard spheres to 0.8 for soft spheres, and A varying within 30%.

The diffusion model presented in this Letter is conceptually distinct from these earlier results in recognizing that (1) liquid dynamics is controlled by the Enskog collisional mechanism, and (2) the rate of structural relaxation is proportional to the phase-space volume, that is, $\propto e^s$. This leads to $D = C\sigma^2\Gamma e^s$, where σ is the atomic diameter, Γ is the collision frequency and C is a universal constant; (1) is corroborated by the finding³ that the Kolmogorov–Sinai entropy in liquids, scaled by Γ , is universally related to s . In contrast to the timescale used in ref. 1, Γ is explicitly structure dependent; therefore, the variations of parameters in ref. 1 can be explained by non-universality of liquid structures. This Letter uses two-particle approximation of s , but a comprehensive test for full s was reported⁴.

An advantage of this relation involving no adjustable parameters is that it makes it possible to test unambiguously the diffusion model it quantifies. Besides liquids, this model holds for solid ionic conductors and quasicrystals. On the other hand, the relation can be used to detect a dynamical transition in supercooled liquids⁵. □

1. Rosenfeld, Y. *Phys. Rev. A* **15**, 2545–2549 (1977).
2. Rosenfeld, Y. *J. Phys. Cond. Matter* **11**, 5415–5427 (1999).
3. Dzugutov, M., Aurell, E. & Vulpiani, A. *Phys. Rev. Lett.* **81**, 1762–1765 (1998).
4. Hoyt, J., Asta, M. & Sadigh, B. *Phys. Rev. Lett.* **85**, 594–597 (2000).
5. Dzugutov, M. *J. Phys. Cond. Matter* **11**, 253–260 (1999).

errata

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Nature **409**, 860–921 (2001).

In this Article, the last name of Joseph Szustakowski was misspelled in the author list under the section Genome Analysis Group. (His name appeared as ‘Joseph Szustakowki’). □

Supplementary information is available from Nature’s World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank the members of the Robbins’ laboratory and A. J. Capobianco, Y. Sanchez, T. Doetschman, L. A. Woollett, S. M. Bell, X. Lin and K. E. Yutzey for discussions. D.J.R. is a recipient of a Burroughs Wellcome Career Development Award. This work was supported by Grants from the National Institutes of Health.

Correspondence and requests for materials should be addressed to D.J.R. (e-mail: david.robbins@uc.edu).

Homologues of Twisted gastrulation are extracellular cofactors in antagonism of BMP signalling

Ian C. Scott, Ira L. Blitz, William N. Pappano, Sarah A. Maas, Ken W. Y. Cho & Daniel S. Greenspan

Nature **410**, 475–478 (2001).

Figures 2 and 3g contained a number of errors. The correct figures are reproduced below. □

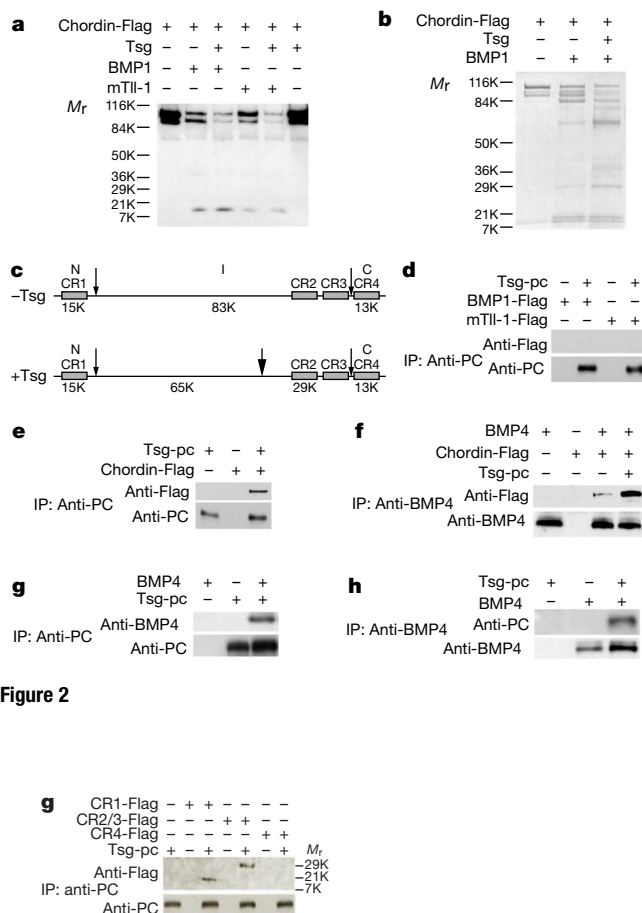


Figure 2

Figure 3g

Liquid refreshment

Automation and a proliferation of channels are *de rigueur* in liquid handling.



Finnpipettes, now with 16 channels.

Finnpipette Digital/BioControl

Thermo Life Sciences

www.thermo-lifesciences.co.uk

Sixteen channels — suitable for 384 wells

The Finnpipette Digital and BioControl ranges now feature 16-channel models, for volumes ranging from 5 to 50 μ l. All the Digital pipettes are fully autoclavable, allowing complete sterility if required. The BioControl electronic pipette can be set in stepper or pipette mode, and has three pipetting speeds. The tip cone is fully autoclavable. The basic pipette can act as a complete liquid handling system, with one universal handle and 13 interchangeable tip cone modules.

Impact/Impact 2

Matrix Technologies www.matrixtechcorp.com

More channels for 384 wells

The new 16-channel Impact electronic pipettor makes it possible to transfer an entire row of a 384-well plate in one go. Available with Impact or Impact 2 software and three different volume ranges, these new pipettors feature a lithium ion battery for long life and rapid recharge rates.

PrepTip

Harvard Bioscience

www.harvardapparatus.com

Sample prep pipette tips

PrepTip, available in sizes from 10 to 1,000 μ l has been developed for cleaning and concentrating proteins, peptides, DNA and other biological samples prior to further analysis. PrepTip features a coating of resins or ligands on

the interior walls of the tip. Since the tip is not packed the sample can flow freely through the opening, making sample preparation easy and suitable for automation. The tips can also be custom coated with materials of your choice. The addition of specific affinity PrepTips (such as avidin-coated tips) expands the range of PrepTip applications still further. Sample clean-up prior to MS-TOF is the most popular application for PrepTip.

AQUAmax 1536

Molecular Devices Corporation

www.moldev.com

Dispense with alacrity

Targeted at high-throughput screening applications, the AQUAmax 1536 dispenser is designed to accelerate assay miniaturization. It uses only small amounts of costly reagent per assay, in volumes of 0.5–10 μ l, reducing assay development costs. The system precisely dispenses reagent quickly and consistently across the microplate for a typical throughput of 30 seconds per plate. Included in the system is Windows-compatible software which allows the user to customize individualized protocols and program the system to accommodate multiple variations of 1,536 microplates.

SciClone ALH

Zymark Corporation

www.zymark.com

Combined function system

The SciClone ALH (Advanced Liquid Handling) Workstation is designed to perform a range of functions that previously required several separate devices, supplying 8-channel independent Z-axis probes (Z-8), 'on-the-fly' interchangeable 96 and 384 arrays, bulk dispensing capabilities and simple integration with other devices. The Z-8 probes permit individual access to test tubes and 96- or 384-well microtitre plates, allowing complete automation of inter-plate and intra-plate

formatting for applications such as hit picking, dilutions and tube-to-microtitre plate transfers. The high-density arrays allow inter-plate formatting for up to 1,536 well plates, making the system particularly suited to many nucleic acid preparations and plate replications.

Pressmatic PP

Bibby Sterilin

www.bibby-sterilin.com

Two sizes, for all bottles

The latest addition to the Pressmatic range, these dispensers are based on a polypropylene valve block with glass piston and barrel. The Pressmatic PP is suitable for use with any liquids compatible with polypropylene, which makes it suitable for routine dispensing of most aqueous, non-aggressive solutions such as buffers, polar solvents, dilute acids and alkalis and biological solutions. Dispensing volume is simple to set using a graduated rotating ring, and the unit is available in 10 and 60 ml sizes, with adapters for



Pressmatic PP: have you got the bottle?

ADVERTISEMENT

Website for Bio-processing Specialists

Website "www.merck.de/bioprocessing" provides a reference source for all those involved with down-stream processing of bio-molecules in the Pharmaceutical Industry. The site not only provides information on the essential chemicals and enzymes for bio-processing but also covers both standard and innovative separation techniques.

Bio-chromatographers can find useful information not only on classical silica based products but also the unique tentacle ion-exchange resins which are used for many applications from removal of viruses to purification of antibodies and DNA. In addition an entire catalogue of products for down-stream bio-processing can be easily downloaded from the site. www.merck.de/bioprocessing

Merck KGaA
Darmstadt · Germany



most commonly used laboratory bottles. A clear window allows the user to check the cylinder for air bubbles.

711 Liquino/700 Dosino

Metrohm www.metrohm.co.uk

Automating for routine preparations

The combination of the 711 Liquino with the 700 Dosino dispensers is offered as a 'work horse' for tasks such as standard preparation for HPLC. The 711 Liquino controls up to four 700 Dosinos with dosing units of 2, 5, 10, 20 or 50 ml. The burette cylinder is available in glass or ETFE. Built-in features include storage and recall of dosing routines, automatic preparation of solutions with defined content, dosing with temperature monitoring and fully automated sampling.

Biomek FX

Beckman Coulter www.beckmancoulter.com

A 384-channel disposable tip pipetting head

The new 384-channel pipetting head for the Biomek FX liquid handling system is easily interchangeable with the 96-channel head to enable rapid interchange between liquid handling methodologies, from tubes to 1,536-well plates. The Biomek FX features Windows NT-based software for creating validated methods, and add-on software wizards for specific applications such as plate replication, nucleic acid preparation and hit picking.

NSOM/AFM LC100

Nanonics Imaging www.nanonics.co.il

Full integration with upright/inverted/dual optical microscopes

This new liquid cell is an accessory for the Nanonics NSOM/AFM 100 Confocal Microscope system. With the liquid cell in place, the microscope's high power objectives can be placed above and below the aperture in the centre of the microscope system. The cell consists of a bath, in which the sample sits, and a tip mount, suitable for Nanonics cantilevered optical fibres or standard AFM silicon cantilevers. It permits all forms of AFM including contact, non-contact and intermittent contact in a liquid environment. The design of the system ensures that the optical axis remains free from the top and the bottom of the liquid cell.

PlateTrak

Packard Bioscience

www.packardbioscience.com; www.ccspackard.com

Take notes on CCS Packard pipettors

Three new Application Notes are available from CCS Packard, Inc. Two of the notes focus on 96- and 384- channel pipetting and the third focuses on cell-based assays using the PlateTrak automated microplate processing system. Pipetting precision data reflect optimized procedures and conditions as detailed

in each Application Note. For copies, call 1-800-323-1891 (US only) or 203-639-2598, or visit one of the company's web sites.

Quad-Z 215

Gilson

www.gilson.com

Multiple probe liquid handler/injector

The Quad-Z has four independently operated probes with variable horizontal spacing (9–18 mm) allowing access to virtually any tube, vial or microplate configuration. The independent operation enables each probe to move on the z-axis to access single samples from the large bed of the Quad-Z 215. An advanced liquid-level detection system on each probe minimizes carryover ensuring accurate and reproducible results. The optional 849 Multiple Injection Module features four injection valves, providing injection capabilities onto HPLC analysis systems utilizing the Quad-Z 215.

BRC2501

Biohit

www.biohit.com

Integrated liquid handling

The BRC2501 is an electronic single channel 250-µl pipetting module designed to offer high performance and flexibility in robotic liquid handling and automated processing. Utilizing Biohit's patented displacement

mechanism, the BRC2501 is claimed to offer accuracy and precision, while avoiding the need for the tubing, solenoids and syringe drives usually associated with such devices. The BRC2501 is controlled by a versatile serial interface (incorporating RS-232 and RS-485) with facilities for integrated level sensing and electronic tip eject.

Dilumat 4

AES Laboratoire www.aeslaboratoire.com

One for the gourmet

Intended primarily for fast and accurate dilution of liquid and solid food samples, the Dilumat 4 allows dilution of a 225 ml sample in just 15 seconds. The latest model is fitted with a motorized dilution arm to minimize manual steps, making it ideal for work in a laminar airflow cabinet. The system is designed for ease of operation — the user simply programs the dilution factor required via the keypad, guided by a straightforward question-and-answer sequence on the display. The Dilumat 4 then weighs the bottle, flask or homogenizer bag containing the sample and automatically calculates and dispenses the exact volume of diluent required. No additional programming is required for serial dilutions.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

OAKTOWN PUBLICATIONS

100 Western Avenue, London W3 7TX
Tel: 020 8749 4555; Fax: 020 8740 7782

Two Seminal books by Shafi U. Ahmed
Fermat's Last Theorem: New Simple Proof
(also Goldbach), ISBN 094 8331 100, and
We Are Alone: The Limited Universe
(Proves speed of light is immeasurable)
ISBN 094 8331 151. Price £10 each plus
P&P £1(UK), £3(abroad).

Order through Booksellers or from us. Major
Credit/Debit Cards accepted. Key texts free on:
www.oaktownpublications.com

Ubiquitin research?

~ polyubiquitin chains ~
~ Ub₄ ~ Ub⁺¹ ~ Ub₅⁺¹ ~
~ Ub[K⁴⁸R] ~ Me-Ub ~
~ biotinyl-Ub ~ His₆-Ub ~
~ Ub-H ~ Ub-AMC ~

www.proteasome.com

AFFINITY Research Products Ltd.
Mamhead Castle, Exeter, EX6 8HD, U.K.
T. +44(0) 1626 891010 – F +44(0) 1626 891090
– Email info@proteasome.com –

BACs • YACs • PACs cDNAs

We provide clones and colony membranes
for a number of human, mouse and rat
libraries as well as a custom screening
service for the libraries.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

CUSTOMISED

immunoassays for
research and production

BioGenes
Gesellschaft für Biopolymere mbH

Internet www.biogenes.de
Berlin • Germany
Phone +49 (0)30-65 76 23 96

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Welss (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Sille Opstrup (4994)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren
Tel + 34 91 3913835
Fax + 34 91 319 0091
e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Shinjuku-Ku,
Tokyo 162
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@nature.com
Japan Manager: Kate Cowan

Positive vibes for biophysics

A recent meeting with ten graduate students and fellows in Rockefeller University's Center for Physics and Biology provided me with a glimpse of the dynamics and demographics behind that multidisciplinary field. But it remains unclear whether that microcosm provides representation or aberration.

First, the demographics. Of the ten, only one came from the United States. The group, selected by admittedly unscientific means (they were available at lunchtime during a recent visit), also contained no female scientists. That roughly meshes with data that show fewer US-born physicists compared with the early 1990s, as well as a continuing dearth of women in the field.

Now, the dynamics. Almost every one of the ten had a different route into the department — one of the first formal biophysics departments in the country. Some came from a maths background, some from theoretical physics, others from condensed-matter physics or statistical mechanics.

But their reasons and aspirations were in sync. All wanted to investigate tangible systems, with real-work applications, rather than the intangibles of high-energy physics. All aspired to academic rather than industrial jobs, even though their skills in statistical modelling are increasingly in demand by both drugs and software industries. And all felt positive about their prospects.

It's fair to say they have good reason to feel this way. The problems they are working on — including using maths, statistics and computational modelling to predict phenomena such as gene regulation — are valuable to all sectors. Those problems are too big and too complicated for biologists to tackle alone. Researchers from the broader scientific universe will probably find plenty of opportunity to grapple with them.

Paul Smaglik
Naturejobs editor



Contents

MOVERS

Husband and wife get closer together; researchers seek breakthrough in breast cancer; and more

[Back page](#)

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

LIFE SCIENCES

Husband-and-wife structural biologists **Jamie Cate** and **Jennifer Doudna** will take up posts at the University of California, Berkeley this year. The moves will allow them to live and work in the same city; their labs are currently a two-hour drive apart. **Cate**, an assistant professor at the Massachusetts Institute of Technology, is moving to California in July. **Doudna**, who has a larger lab at Yale — which will take longer to move — will join him later. “We’re going to be bi-coastal for a little while,” Doudna says. At Berkeley both will maintain separate labs and projects. Cate works on protein-synthesis mechanisms, including the structure of the ribosome at different states. Doudna studies catalytic RNAs and RNAs that interact with the ribosome. In addition to allowing them to work closer to each other, the move also means they will be closer to synchrotron beamlines, needed for X-ray crystallography elucidation of their molecules of interest.

British immunologist **Michael Owen** is this month leaving the Imperial Cancer Research Foundation (ICRF) for a position with GlaxoSmithKline (GSK). Owen, who joined the ICRF in 1979, has spent much of his career characterizing T cells, a key part of the immune system. The biggest change, he says, will be moving from studying “a very small problem in a great deal of detail” to analysing broader, more open-ended problems. As senior vice-president and head of technology development with GSK, Owen will direct a multidisciplinary team of scientists in finding methods to weed out drug leads that look promising *in vitro* but may not work in humans. He plans to look at biotech companies, universities and GSK for new technology that could make drug screening more efficient.

Two new researchers have joined the Breakthrough Breast Cancer Research Centre in London — **Sunil Lakhani** as professor of breast cancer pathology and **Clare Isacke** as professor of cell biology. **Lakhani** was previously consultant and reader in pathology at University College Hospital in London. His research interests are in early lesions, familial breast cancer and myoepithelial tumours. **Isacke** was previously professor of cell biology at Imperial College, London. Her interests are in cell–cell and cell–matrix interactions as applied to the process of metastasis. The centre, located within the Chester Beatty Labs of the Institute of

Cancer Research/Royal Marsden Hospital Trust complex, opened at the end of 1999. Its focus is to foster interactions between basic biologists and clinicians. Eventually, the facility will encompass 100 staff, of which 70 have so far been appointed.

PHILANTHROPY

Helene Gayle, director of the National Center for HIV, STD, and TB Prevention, which is run by the US Centers for Disease Control and Prevention (CDC), will take up a newly created position as senior advisor for HIV/AIDS with the Bill and Melinda Gates Foundation in September. In her six years as director of the CDC’s main centre dealing with HIV/AIDS, other sexually transmitted diseases and tuberculosis prevention, Gayle has expanded community-based HIV-prevention activities, particularly for minority and underserved communities. She currently oversees a multidisciplinary staff of more than 1,400 employees and a budget of around \$1 billion. At the Seattle-based foundation, she will oversee a \$300 million budget for AIDS treatment and prevention and will work to create public–private research and treatment links.

INFORMATION SCIENCE

Lorcan Dempsey, director of the Distributed National Electronic Resource at King’s College, London, will move to Dublin, Ohio, to join the Online Computer Library Center (OCLC) on 16 July as vice-president of research. The Distributed National Electronic Resource is a UK initiative to organize Britain’s higher-education information infrastructure. The OCLC is a non-profit organization that provides computer-based cataloguing, reference, resource sharing and preservation services to 39,000 libraries in 76 countries and territories.

SPACE SCIENCE

Jonathan Ormes has been named director of space sciences at NASA’s Goddard Space Flight Center in Greenbelt, Maryland. Ormes will be responsible for research ranging from flight-experiment development to data analysis. After Ormes received his PhD from Stanford in 1967, he joined NASA under a US National Academy of Sciences resident research associateship programme — a programme he would like to revive.



Jennifer Doudna



Jamie Cate



Clare Isacke



Sunil Lakhani

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.